TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Computer Science Division of Bioinformatics, Biotechnology Center TU Dresden

UNDERSTANDING THE STRUCTURAL AND FUNCTIONAL IMPORTANCE OF EARLY FOLDING RESIDUES IN PROTEIN STRUCTURES

Sebastian Bittrich

Born on: January 19, 1990 in Karl-Marx-Stadt, now called Chemnitz

DISSERTATION

to achieve the academic degree

DOCTOR RERUM NATURALIUM (DR. RER. NAT.)

Referee

Prof. Dr. Peter Stadler

Advisor

Prof. Dr. Dirk Labudde

Supervisor

Prof. Dr. Michael Schroeder

Submitted on: November 20, 2018

Defended on: January 9, 2019

LIST OF PUBLICATIONS USED IN THE THESIS

1.

Bittrich, S., Schroeder, M., & Labudde, D. (2018). Characterizing the relation of functional and Early Folding Residues in protein structures using the example of aminoacyl-tRNA synthetases. *PloS one*. 13(10), e206369.

Contribution: SB conceived and designed this study, analyzed the data, and drafted the paper.

Thesis: In Chapter 5 the findings of this paper are presented and discussed in detail. The proposed set of features is utilized in Chapter 6. The case study on aminoacyl-tRNA synthetases is part of Chapter 7.

2.

Bittrich, S.☯, Kaden, M.☯, Leberecht, C., Kaiser, F., Villmann, T., & Labudde, D. (2018). Application of an Interpretable Classification Model on Early Folding Residues during Protein Folding. *BioData Mining*, in press.

Contribution: SB co-designed this study, implemented the core algorithm, and guided the writing process.

Thesis: Chapter 6 is based on this paper, wherein a novel classification model for early folding residues is presented.

3.

Bittrich, S., Heinke, F., & Labudde, D. (2016). eQuant – A Server for Fast Protein Model Quality Assessment by Integrating High-Dimensional Data and Machine Learning. In *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery* (pp. 419-433). Springer, Cham.

Contribution: SB designed this study, implemented the functionality of the server, and guided the writing process.

Thesis: Energy Profiles are used to represent protein structures in Chapter 5 and Chapter 6. Section 5.1.3 presents the Energy Profiling approach in detail.

4.

Kaiser, F.☯, Bittrich, S.☯, Salentin, S., Leberecht, C., Haupt, V. J., Krautwurst, S., Schroeder, M., & Labudde, D. (2018). Backbone brackets and arginine tweezers delineate class I and class II aminoacyl tRNA synthetases. *PLoS computational biology*, 14(4), e1006101.

Contribution: SB co-designed this study, contributed to the definition of the dataset, co-analyzed the data, and contributed to the writing process.

Thesis: In Chapter 7 this enzyme superfamily is presented: two structural motifs were identified which are discussed in an evolutionary context.

☯ – authors contributed equally

ADDITIONAL PUBLICATIONS

5.

Kaiser, F., Eisold, A., Bittrich, S., & Labudde, D. (2015). Fit3D: a web application for highly accurate screening of spatial residue patterns in protein structure data. *Bioinformatics*, 32(5), 792-794.

6.

Heinke, F., Bittrich, S., Kaiser, F., & Labudde, D. (2016). SequenceCEROSENE: a computational method and web server to visualize spatial residue neighborhoods at the sequence level. *BioData mining*, 9(1), 6.

7.

Szendy, M., Kalkhof, S., Bittrich, S., Kaiser, F., Leberecht, C., Labudde, D., & Noll, M. (2018). Structural change in GadD2 of Listeria monocytogenes field isolates supports nisin resistance. Under review by *International Journal of Food Microbiology*.

ACKNOWLEDGEMENTS

I am grateful for the support of my supervisors Michael Schroeder and Dirk Labudde. Dirk established the foundation of this thesis and gave me the opportunity to pursue academic work after my graduation. Michael shared his highly intriguing view on what makes a good publication, dissertation, and scientist. The supervision of both shaped this thesis.

Furthermore, Michael constitutes the connection to Sebastian Salentin and Joachim Haupt. Both taught me how to manage projects efficiently and forged joyful collaborations especially on the topic of aminoacyl-tRNA synthetases. Florian Heinke supervised my bachelor and master project and, thus, influenced me deeply as a scientist. It is his merit that I am working on my PhD thesis now. I greatly appreciate scientific discussions, constructive criticism, nice work environment, and friendship with Florian Kaiser and Christoph Leberecht. In particular, all of us are connected by the project on aminoacyl-tRNA synthetases which became an integral part of this thesis. Peter Wills first intrigued us about the topic in general and governed the direction our collaboration took ever since. He too shaped the narrative of this thesis.

All my co-authors also supported me substantially and influenced the character of the underlying publications. Sarah Krautwurst, Marika Kaden, Lauren Adelmann, Hanna Sieweerts, Steffen Grunert, Stefan Schildbach, Alexander Eisold, Thomas Villmann, and Marcel Scheuche contributed to various projects, provided valuable feedback, assisted behind the scenes, and/or proofread manuscripts.

Also, I appreciated all conversations with people who shared their experiences and point of view in academia, industry, and publishing. Gratitude is owed to my family and friends who supported me during this period, motivated me, and made this an enjoyable endeavor. You are the basis for me to do scientific work at all. Specifically, I would like to thank Annemarie and Rick for proofreading this thesis.

And it all comes down to you
Well you know that it does
And lightning strikes, maybe once, maybe twice

Fleetwood Mac, Gypsy

TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Computer Science Division of Bioinformatics, Biotechnology Center TU Dresden

ABSTRACT

Proteins adopt three-dimensional structures which serve as a starting point to understand protein function and their evolutionary ancestry. It is unclear how proteins fold *in vivo* and how this process can be recreated *in silico* in order to predict protein structure from sequence. Contact maps are a possibility to describe whether two residues are in spatial proximity and structures can be derived from this simplified representation. Coevolution or supervised machine learning techniques can compute contact maps from sequence: however, these approaches only predict sparse subsets of the actual contact map. It is shown that the composition of these subsets substantially influences the achievable reconstruction quality because most information in a contact map is redundant. No strategy was proposed which identifies unique contacts for which no redundant backup exists.

The StructureDistiller algorithm quantifies the structural relevance of individual contacts and identifies crucial contacts in protein structures. It is demonstrated that using this information the reconstruction performance on a sparse subset of a contact map is increased by 0.4 Å, which constitutes a substantial performance gain. The set of the most relevant contacts in a map is also more resilient to false positively predicted contacts: up to 6% of false positives are compensated before reconstruction quality matches a naive selection of contacts without any false positive contacts. This information is invaluable for the training to new structure prediction methods and provides insights into how robustness and information content of contact maps can be improved.

In literature, the relevance of two types of residues for *in vivo* folding has been described. Early folding residues initiate the folding process, whereas highly stable residues prevent spontaneous unfolding events. The structural relevance score proposed by this thesis is employed to characterize both types of residues. Early folding residues form pivotal secondary structure elements, but their structural relevance is average. In contrast, highly stable residues exhibit significantly increased structural relevance. This implies that residues crucial for the folding process are not relevant for structural integrity and vice versa. The position of early folding residues is preserved over the course of evolution as demonstrated for two ancient regions shared by all aminoacyl-tRNA synthetases. One arrangement of folding initiation sites resembles an ancient and widely distributed structural packing motif and captures how reverberations of the earliest periods of life can still be observed in contemporary protein structures.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

aaRS aminoacyl-tRNA synthetases

ABD anticodon binding domain

auROC area under the receiver-operating characteristic

CATH class, architecture, topology, homologous superfamily

CCM classification correlation matrix

CM contact map

CP1 connecting peptide

DCA direct coupling analysis

DI direct information

DM distance map

DNA deoxyribonucleic acid

DSSP dictionary of protein secondary structure

EC number of enzyme in Enzyme Commission's system

EFR early folding residues

eProS energy profile suite

GLVQ generalized learning vector quantization

GMLVQ generalized matrix learning vector quantization

HDX hydrogen-deuterium exchange

HSR highly stable residues

ID insertion domain

LFR late folding residues

LVQ learning vector quantization

MI mutual information

MS mass spectrometry

MSA multiple sequence alignment

NB naive Bayes

NMR nuclear magnetic resonance

PDB protein data bank

PLIP protein-ligand interaction profiler

RASA relative accessible surface area

RF random forest

RMSD root-mean-square deviation

RNA ribonucleic acid

ROC receiver-operating characteristic

SCOP structural classification of proteins

SGA stochastic gradient ascent

SIFTS structure integration with function, taxonomy and sequences resource

SML supervised machine learning

SOM self-organizing map

SVM support vector machine

TM-score template modeling score

UR unstable residues

Weka Waikato environment for knowledge analysis

x-ray X-ray crystallography

1. INTRODUCTION

1.1. MOTIVATION

Proteins Are Ambivalent Protagonists Nature fascinates by a seemingly effortless elegance of its processes as well as solutions to problems believed unsolvable. Especially proteins make a case for this statement as they are entities involved in virtually all aspects of life. They may even be referred to as the essence of life. Proteins are chains of only 20 distinct amino acids, yet they implement the astonishing diversity of phenotypes which life encompasses. Proteins adopt three-dimensional structures which seem fragile yet prove reliable whenever needed. The implementation of some functions such as protein biosynthesis did not change drastically since the emergence of life, yet organisms can develop resistance to antibiotics synthesized only decades ago. Conservatism and progressivism seem carefully balanced in the evolutionary process. Interestingly, well-adapted species with large population sizes have little incentive to risk their position and evolve slowly. In contrast, small populations which battle for their existence will fixate genetic change more easily and may evolve strategies to cope with harsh conditions [3]. Whether a species propers or is decimated is a story ultimately told by their respective set of proteins.

Protein Function Matters the Most A gene encodes the amino acid sequence which will constitute a specific protein. In a process called protein folding, this chain of amino acids adopts a three-dimensional structure. Ultimately, this structure can implement functions such as an enzymatic reaction or the propagation of molecules or signals. Over the course of evolution, the function of specific proteins was changed, reshaped, or refined. For this to happen, the underlying gene has to change, because only at this level variation can manifest and be passed down to ancestors. What truly matters is this function of a protein (Figure 1.1). It is of subordinate importance which structure harbors this function and what sequence encodes that structure. The primary requirement is that the function is intact, regardless of the means used to achieve this goal. This is referred to as "functionalist principle" and implies that function is evolutionarily more conserved than structure, which in turn is more conserved than sequence [4]. Still, protein sequence, structure, and function are strongly entangled.

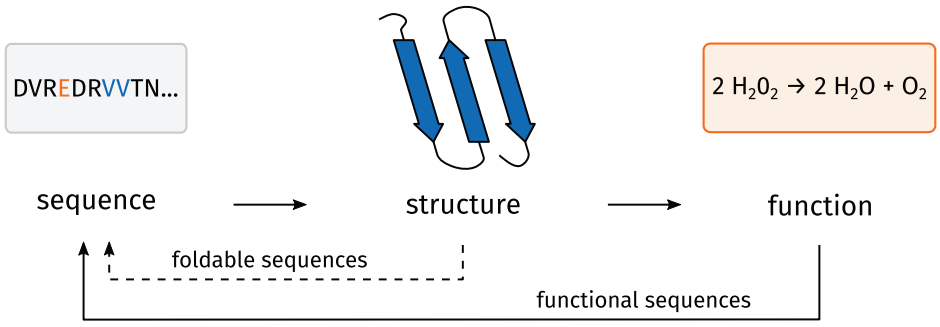

Figure 1.1.: Evolutionary Pressure Acting on Proteins

An amino acid sequence describes the composition of a protein. During protein folding, this chain of residues folds into a three-dimensional structure. This defined structure usually provides the scaffold to implement function (such as an enzymatic reaction, here the example of a catalase degrading hydrogen peroxide). Evolutionary change can only manifest at sequence level. According to the functionalist principle, the primary evolutionary pressure acts on protein function. To a lesser degree, foldable sequences which will fold reliably are selected as well. This implies that function is more conserved than structure, which in turn is more conserved than sequence.

Reductionism Can Be Observed in Nature Proteins are highly complex entities which consist of thousands of atoms. In some cases, slight disturbance of their relative arrangement can lead to diseases such as Alzheimer's, Parkinson's, or amyotrophic lateral sclerosis. The number of possibilities to arrange atoms in the three-dimensional space is incomprehensible. The same is true for the possible combinations of amino acids which may constitute a protein. However, the amount of sequences and structures ascertained over the course of evolution is surprisingly small [5]. A limited set of sequences and structures can implement virtually unlimited function which can e.g. be observed in antibodies, a group of proteins which recognize and bind other molecules. Oftentimes, protein function is realized by a handful of residues. Aspartate, histidine, and serine form the catalytic triad in some proteases. Two residues bind the ATP ligand in aminoacyl-tRNA synthetases. It is not the arrangement of all thousands of atoms that matters, but rather the arrangement of residues in a spatially limited active site. The rest of the protein ensures the correct orientation of this binding site and may modulate its properties such as the affinity for some ligands. The overwhelming complexity of protein structures can be reduced. If protein function can be broken down to a small number of residues, the same may be true for the structure of proteins. Occasionally, key residues for structural integrity (e.g. residues which prevent misfolding) are identified in protein structures. However, such results cannot be directly compared to the findings of other studies as their annotation is context-specific. Similar reductionism can be observed in nature as sequences, structures, and functional mechanisms tend to reoccur. Established strategies are reused rather than reinvented. Scientists use reductionistic approaches as well to describe protein structures: contact maps, residue graphs, coarse-grained energy models, or structural motifs. Only by understanding the functional and structural building blocks of nature (Figure 1.2) it is possible to improve the classification of novel protein sequence as well as structure prediction [6, 7].

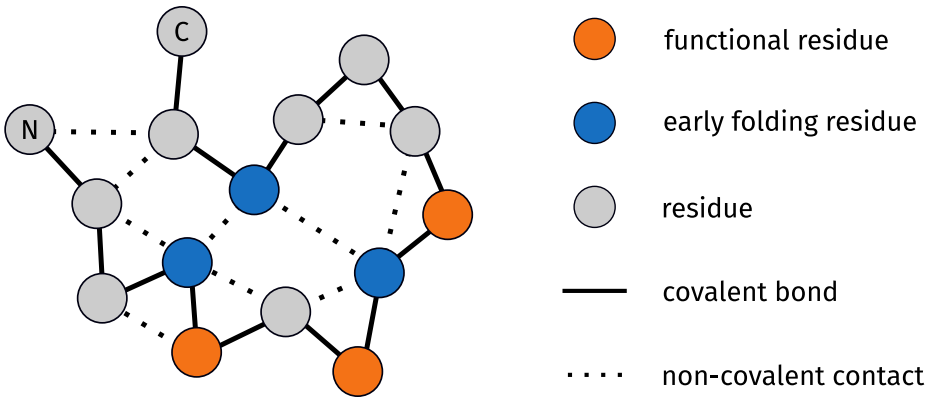

Figure 1.2.: Functional and Structural Reductionism in Proteins

Proteins are chains of covalently bound amino acids (shown here as circles). Its N- and C-termini are given. Some residues implement protein function (orange), a set of different residues tends to fold early during protein folding and may be of importance for the structural integrity (blue). A large number of residues has no direct influence on either aspect. Interestingly, crucial aspects of proteins are commonly realized by a small number of key residues. Covalent bonds impose strong constraints on the spatial location of both neighbors of a residue. Non-covalent contacts can be formed freely between distant residues (so-called tertiary contacts). It is unclear what causes a non-covalent contact to form in a protein structure, while other possibilities are not realized.

On the Relation of Function and Structure It is unclear what causes a certain sequence to adopt a defined fold. Even seemingly unrelated sequences may result in a similar fold [5]. Proteins occur within the crowded environment of a cell and are influenced by factors such

as physiological conditions, ligands, and chaperones. For a long time, scientists have been trying to establish a theory of everything [3] which relates the enigmatic and contradictory aspects of proteins. This thesis cannot provide that. However, to converge further on the problem the structural and functional importance of particular residues in protein structures is discussed. Thanks to new experimental data, this connection can be analyzed for the first time. A standardized set of folding initiation sites (so-called early folding residues) has been published in 2016 and provides general data on folding characteristics of a diverse set of proteins [8]. This dataset also contains information on highly stable residues which prevent spontaneous unfolding of proteins. The scientific community has not yet analyzed this promising resource thoroughly.

Early folding residues may be a valuable resource to understand the folding process in more detail as they provide information on folding intermediates. The sequence resembles the starting point of the process while the native structure is the end point; what happens during protein folding is enigmatic. The inner workings of protein folding are assumed to be essential to understand the process [9, 10, 11]. Other authors argue that such special aspects can only be correctly interpreted if protein folding in general was understood [12].

1.2. AIM

Considering the previous statements, the objective of this thesis is to:

> Aim
>
> Characterize the relevance of early folding residues in protein structures

The exact importance of early folding residues for the folding process is unclear: is the observed signal the mere consequence of undirected physical chemistry or will early folding residues exhibit distinct properties in the folded structure as well? In previous studies [13], early folding residues have been shown to preferably occur in the hydrophobic core of proteins, embedded in ordered secondary structure elements. Furthermore, early folding residues tend to be the most connected residues in the native structure. While this analysis is a first assessment, a more detailed investigation with more specialized features is still pending. Especially an analysis to other key residues in proteins structures (such as those implementing function) may provide new insights into the connection of sequence, structure, and function. Knowledge of early folding residues and highly stable residues can advance structure prediction routines. Intuitively, there should be some equivalence between *in vivo* and *in silico* folding (Figure 1.3). A contact map is a reduced representation of a protein that captures which residues are in spatial proximity. Contact maps are integral for state-of-the-art structure prediction techniques and may also give insights into the problem of protein folding when related to experimentally determined folding characteristics as provided by early folding residues and highly stable residues. Commonly, protein structures are computed from contact map representation. Yet, they are little understood, especially when it comes to the most influential contacts. Which contacts must be known to yield good structure predictions? Which contacts provide the most information, so that the number of known contacts can be minimized [14, 15]? Most crucially contact maps are sensitive to false positive predictions (i.e. contacts not actually present in the native structure) [16, 17]. A strategy is necessary to interpret contact maps in detail to answer the previous questions.

> Open Question I
>
> How to identify structurally relevant residue contacts in a contact map?

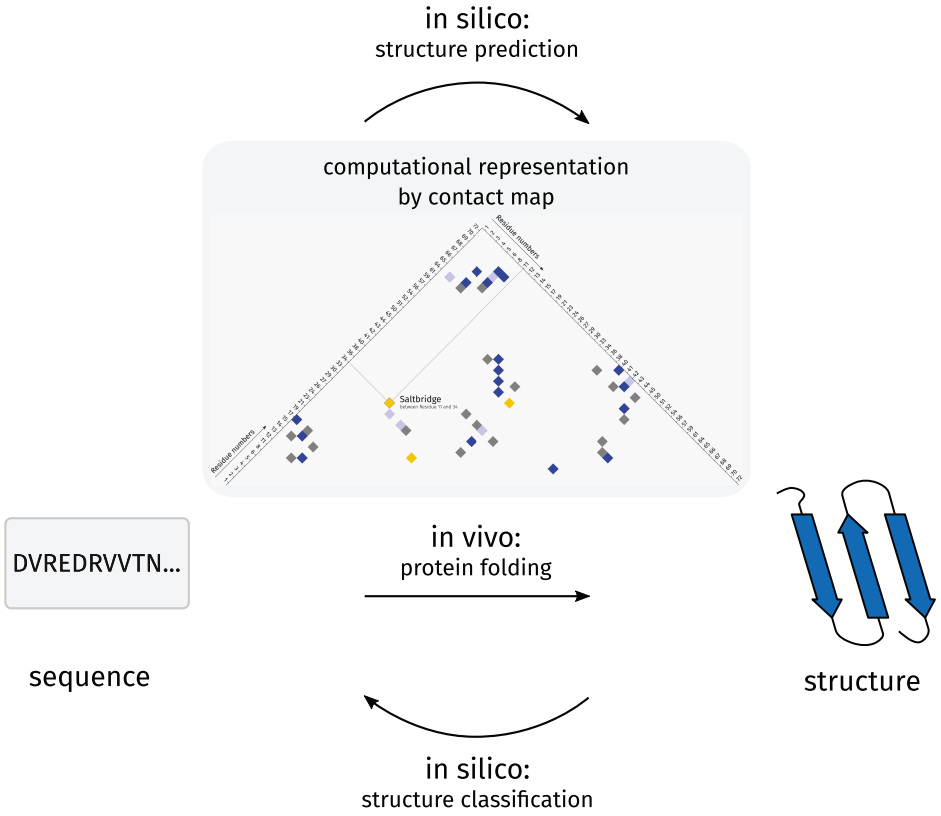

Figure 1.3.: Relation of *In Vivo* Protein Folding and *In Silico* Structure Prediction

The connection of protein sequence and structure remains elusive. *In vivo*, proteins fold into their native structure, *in silico* structure prediction methods aim at recreating this process by simulations and predict the structure that a particular sequence will adopt. Contact maps are potent tools relevant for both processes. A contact map (the central element in pyramid shape) captures which residues are in spatial proximity (indicated by a square); the majority of a contact map is empty because residues are only in spatial proximity to a small number of other residues. Regarding *in vivo* protein folding, contact maps can be used to assess the role of residues with special folding characteristics such as early folding residues and highly stable residues. Furthermore, contact maps may help understanding what the "structural essence" of a protein is: the minimal set of contacts which captures its structure. This thesis will address these questions although a more fine-grained interpretation of contact maps has to be devised to achieve this goal.

Contact maps can be used to create all-atom reconstructions which resemble the native structure of a protein. A fine-grained interpretation of a contact map can be achieved by quantifying the improvement of this reconstruction when a particular contact is known. Throughout the thesis, the term "structural relevance" will be used to refer to the performance gain constituted by the knowledge of a particular contact. It is difficult to assess the structural relevance of individual contacts, because every individual contact depends on all other contacts of a protein structure. The major challenge is the combinatorial complexity: if as little as 50 contacts are present in a small protein structure, the number of combinations to assess (50!) would have 65 decimal digits. Some simplifying strategy has to be devised to disentangle individual contacts from the rest of the protein while attributing to its context-specificity. It should be possible to propose a strategy which can quantify the structural relevance of individual residues and residue-residue contacts in protein structures.

Early folding residues indicate that protein folding is deterministic in that the same residues initiate the folding process and that the order in which certain structural elements are established does not vary when the protein folding process is observed multiple times [18, 12,

19]. Also, early folding residues have been related to sequence fragments which exhibit a rigid backbone [20] and therefore have an increased propensity to form ordered secondary structure elements [13, 11]. It is unclear which intramolecular interactions are the most important ones during protein folding. Prominent candidates are hydrogen bonds, which are primarily observed in the internal stabilization of secondary structure elements, and hydrophobic interactions, which often furnish tertiary contacts (i.e. more than five positions apart at sequence level) between secondary structure elements [21]. Knowledge of the most relevant contacts in a protein structure may advance structure prediction techniques by increasing the performance on sparse or erroneous data. In general, a new strategy to identify the most relevant contacts would also make the information present in contact maps more tangible: for now the influence of individual contacts cannot be quantified and analyzed.

Residues important for the folding process have also been described to be non-functional residues [22]. Modularity is commonly observed in nature [23, 24]. It may be advantageous if function and structure are separated in proteins, so that function (e.g. the ligand recognized by a binding site) can change without compromising the fold of a protein. To investigate this question in detail, the evolutionary history of a diverse protein superfamily has to be studied. Aminoacyl-tRNA synthetases are present in all organisms and are some of most ancient proteins, potentially dating back 4 billion years. Therefore, they are prime candidates to study their evolutionary history and relate early folding residues and functional residues in these structures.

Open Question II

How are early folding residues affected by evolution?

In literature, functional residues are often highly conserved positions because only a narrow set of amino acids can realize most functions [25]. Folding initiation sites tend to be more conserved than other residues too [22]. No study describes how early folding residues are conserved in a diverse set of proteins due to sparseness of the required data. However, a predictor for early folding residues has been published [11] which can be used to approach the relation of structurally and functionally relevant residues. In particular, an open question is how folding initiation sites are conserved over the course of evolution. Does their position change for homologues? Are they more conserved than functional residues? This knowledge is e.g. required to understand how virus proteins escape recognition by immune systems [26].

1.3. OUTLINE

Background In Chapter 2, folding theories and underlying experimental techniques are presented. In this context, evolutionary aspects of proteins are discussed. Chapter 3 provides details on contact maps as simplified representations of protein structure and their relevance for structure prediction methods.

Results and Discussion Structurally and functionally relevant residues were assessed in terms of the relevance for structural integrity in Chapter 4 by the StructureDistiller algorithm. In Chapter 5 the direct relation of structurally and functionally relevant residues is studied. A detailed analysis on aminoacyl-tRNA synthetases substantiates these findings. A machine learning algorithm was implemented and applied on the problem of early folding residues in Chapter 6. Chapter 7 presents background on aminoacyl-tRNA synthetases and assesses whether early folding residues are evolutionarily conserved. The conclusion in Chapter 8 will revisit the questions put forward here.

Part I.

BACKGROUND

2. THE PROTEIN FOLDING PROBLEM

In the process of protein folding (Figure 2.1), a denatured protein chain without structure (*D*) adopts a defined, native conformation (*N*). This native state allows most proteins to fulfill their function and thus to be biologically relevant. Native structures are publicly available in the protein data bank (PDB). The denatured state exhibits high free energy and entropy, whereas the native conformation is one of low free energy and high order [27, 28, 18]. The transition state (‡) is located between both states and acts as an energetic barrier [27, 28]:

$$D \rightleftharpoons \ddagger \rightleftharpoons N \tag{2.1}$$

The transient nature of the transition state hinders the initial formation of the native conformation, but also limits spontaneous unfolding [27, 28]. The unfavorable parameters of the transition state may be the consequence of an exposure of many hydrophobic residues to the solvent, which provides an additional incentive to progress toward the native conformation [29]. The difference in free energy achieved by protein folding is relatively small. In consequence, the conformation must realize all potential, favorable interactions as seen e.g. in hydrogen bonds. It is remarkable that most proteins fold autonomously because finding a solution to this problem is difficult [27]. Proteins feature a hydrophobic core where residues are excluded from the surrounding solvent and are stabilized e.g. by tertiary contacts (i.e. residues which are more than five positions apart at sequence level).

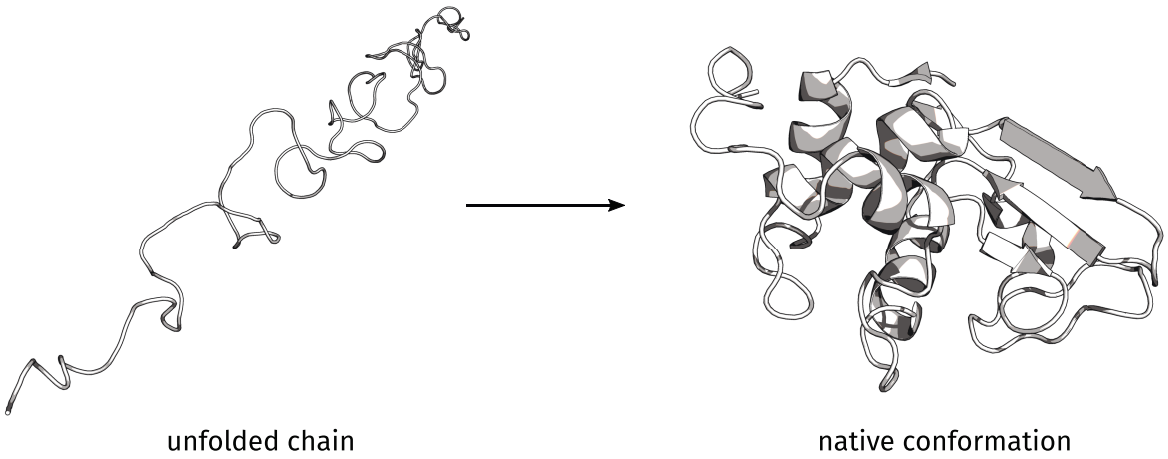

Figure 2.1.: The Process of Protein Folding

Protein folding [27, 28] describes the process in which an unfolded protein chain (left) adopts its corresponding native conformation (right). A change from high entropy and high energy to low entropy and low energy can be observed. Usually, the native structure allows a protein to be biologically active. Remarkably, most proteins can fold autonomously. How so is little understood.

The folding of proteins cannot be reliably modeled or described *in silico*. This is the result of the limited understanding for the process of protein folding in general. Anfinsen et al. demonstrated that proteins can fold autonomously [30]. Later, Levinthal argued that there must be some order to the process. A completely random sampling of all possible conformations of a protein chain would require unfeasibly large time [31]. It is evident that protein folding is an optimization problem: find the most stable spatial arrangement of all atoms for a given protein sequence. This problem seems to be solved by finding the solution to partial optimization problems for fragments and assembling these fragments to yield the global structure while limiting the degrees of freedom [27]. This implies that kinetic rather than stability determines protein folding. Proteins find a good solution fast, but the global optimum may only be found by a more exhaustive search [32]. Zwanzig et al. showed that single contacts cannot reliably stabilize proteins, so there has to be a complex network of

interactions present to formate the native conformation [33]. Commonly, it is assumed that native structures exhibit low free energy (Figure 2.2), although there are cases where non-native conformations such as protein aggregates feature even lower free energy. Small proteins tend to fold fast in a so-called two-state manner, wherein folding intermediates exist but do not accumulate significantly (as described by the first equation). Two-state folding is attributed to small, fast folding proteins with simple or single domain architecture [34].

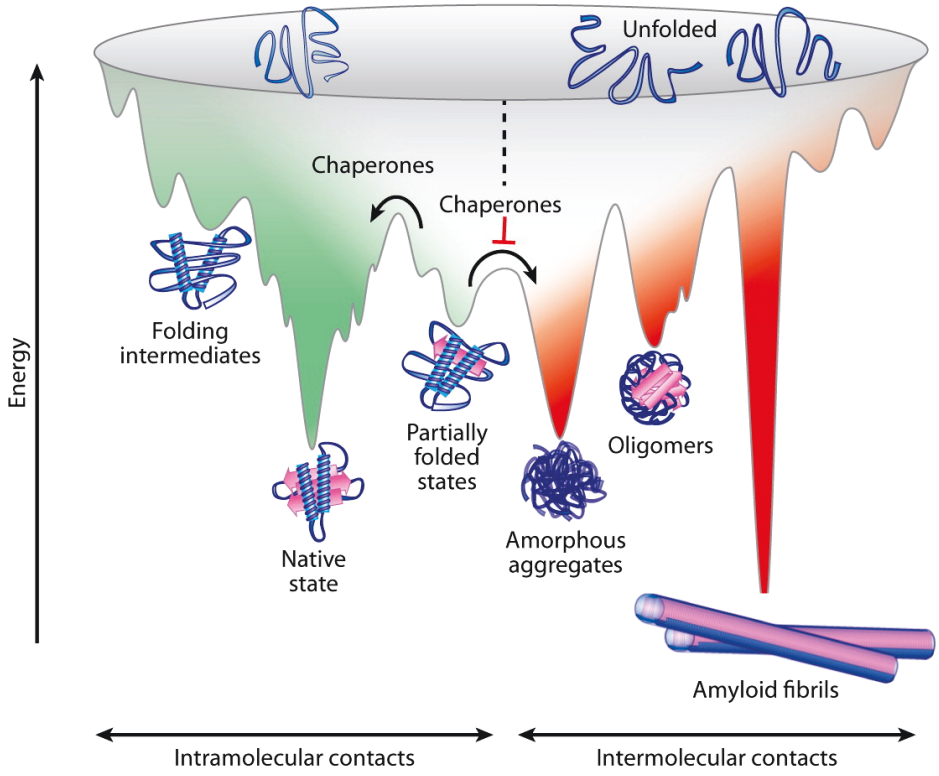

Figure 2.2.: Proteins Strive for Low Energy

Unfolded proteins and folding intermediates exhibit states of relatively high free energy, whereas the native conformation resembles low energy states. Most proteins fold autonomously, while some depend on chaperones to fold correctly. Chaperones can also help preserving the native conformation of proteins or prevent the aggregation of misfolded proteins (wells of the energy landscape colored in red) [1].

2.1. STABILIZING NON-COVALENT CONTACTS IN PROTEIN STRUCTURES

Various types of contacts stabilize macromolecules or interactions between them. Understanding these interactions is the key to converge on the protein folding problem [35]. Similar interactions occur between a protein and potential ligands [36, 37]. Interestingly, these interactions are also a key aspect when differences of meso- and thermostable proteins are studied [38, 39]. An established tool for the detection of both intra- and intermolecular non-covalent contacts is the protein-ligand interaction profiler (PLIP) [37]. Especially, PLIP supports the fine-grained detection of various interaction types. The PLIP program was used for the annotation of such interactions due to direct support by the developers. To my knowledge it is the only tool which handles inter- and intramolecular contacts and provided consistent behavior for this thesis as both aspects were used on various occa-

sions. The physicochemical properties of amino acids determine which contact types they can participate in (Figure 2.3).

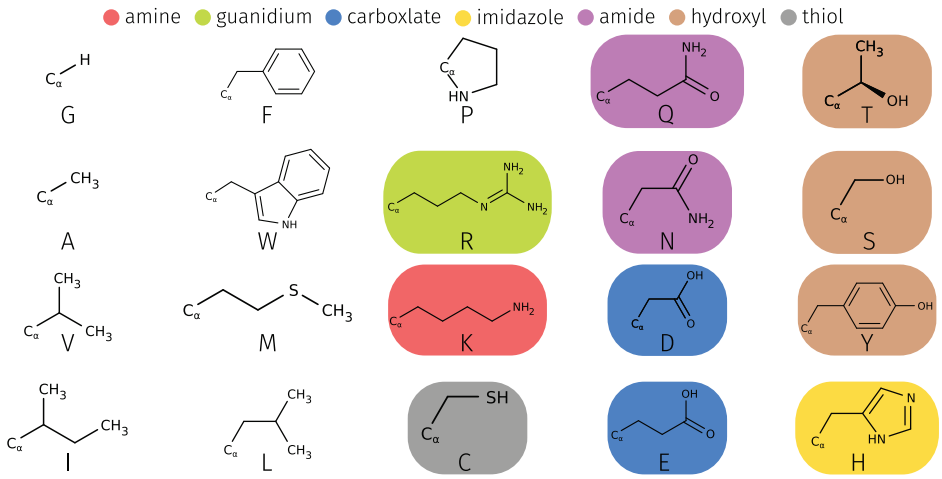

Figure 2.3.: The 20 Canonical Amino Acids

Amino acids are colored by the grouping of Gutteridge et al. [25]: a subset of amino acids is catalytically active because their side chain includes polar groups. Amino acids on the left are more likely structurally relevant or modulate characteristics of the binding site. The physicochemical properties also determine which contacts an amino acid can participate in. Figure adapted from [2].

Hydrogen Bonds Hydrogen bonds [40] are well-known for their role regarding the initiation and stabilization of secondary structure elements. These interactions occur between backbone atoms of the corresponding residues. However, especially at the start and end points of ordered secondary structure elements, this hydrogen bond pattern is not present so that backbone hydrogen bonds can be formed freely to other residues [27]. Some amino acids such as threonine or serine can form hydrogen bonds via their side chain atoms [41]. Moreover, simplistic models [42, 43] suggest that the protein fold ought to be written in the side chains of the corresponding amino acids because the potential to form backbone hydrogen bonds is given at all sequence positions: all amino acids can form hydrogen bonds furnished by their backbone atoms [27]. Hydrogen bonds are considered the most important directed, non-covalent interaction in macromolecules [44]. They are formed between a donor group with a partial positive charge at its hydrogen atom and acceptor group with high electron density [45]. The energetic contribution of hydrogen bonds is context-dependent [41] and it is argued that the unpolar environment in the hydrophobic core of a protein makes hydrogen bonding even more favorable [46, 47]. This also implies that *in silico* modeling of helices may benefit from a dedicated representation tailored for polar environment and one for hydrophobic environment, i.e. the interior of a protein or a cell membrane [48, 49]. Potentially the propensity to form helix-helix interactions changes drastically. Weak hydrogen bonds (C–H· · ·O) contribute to a lesser degree to protein stability [44], with their exact importance for protein folding still being discussed [50, 51, 45, 44]. The formation of hydrogen bonds (and thereby, the folding of a protein) can be prevented by GuHCl [41, 27].

Hydrophobic Interactions Amino acids are considered aliphatic when their side chain features only carbon atoms (ignoring hydrogen atoms). This encompasses alanine, valine, isoleucine, and leucine. Commonly, glycine is also considered aliphatic as its side chain is a hydrogen atom. The larger the corresponding side chain of an aliphatic amino acid, the higher is its hydrophobicity. Valine, isoleucine, and leucine contribute to the formation of the

hydrophobic core and play a substantial role in stabilizing protein structures [52, 53]. This positive effect is attributed to the reduction of the exposed area of hydrophobic side chains to the polar solvent [54]. Hydrophobic interactions are linked to a potential hydrophobic collapse during protein folding. How the hydrophobic core of a protein is established is still in debate [55, 27]. There are cases where the hydrophobic collapse to a molten globule precedes the formation of secondary structure elements [56]. Some of the earliest *in silico* folding models [57, 58, 21] yielded remarkable results by only considering the hydrophobicity of residues.

Aromatic Interactions The importance of aromatic interactions for the stabilization of proteins is well-known [59]. The aromatic amino acids phenylalanine, tyrosine, and tryptophan are usually buried, i.e. they are isolated from the polar solvent and are embedded in the hydrophobic environment of the protein core. These aromatic amino acids share an increased propensity to be occurring together and to form interaction networks of more than two residues. π-stacking interactions in proteins predominantly occur in a perpendicular orientation; this opposes the parallel orientation observed in deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). Aromatic interactions result from attractions of delocalized π-electrons [59]. When histidine is not charged, it forms π-interactions due to its latent aromatic nature and contributes favorably to protein stability [60]. Both hydrophobic and aromatic amino acids have been linked to ordered secondary structure elements [20, 8]. These ordered secondary structure elements seem to be evolutionary conserved [11].

π-Cation Interactions Cationic side chains of lysine or arginine interact favorably with aromatic rings. The positive charge attracts the excess of delocalized electrons above or below aromatic rings and this interaction stabilizes proteins as well as protein-protein interfaces [61, 62, 63]. If histidine residues are positively charged, they can participate in π-cation interactions as cation [60].

Salt Bridges Salt bridges are furnished by residues in spatial proximity which exhibit complementary charges. Thereby, they encounter an attractive force like that between ions in salt [64]. It has been shown that the introduction of salt bridges increases the thermostability of proteins [65] and they have been linked to increased thermostability of particular proteins [38, 39]. When charged groups are buried in the hydrophobic core of a protein, the structure tends to be destabilized. However, salt bridges in a hydrophobic environment are not strictly unfavorable but may also implement specificity to encode the proper fold of a protein [66] because they provide an incentive for defined, charged residues to interact.

2.2. FOLDING THEORIES

Ever since the first crystal structure of a protein [67] has been published, scientists try to understand how proteins adopt a defined three-dimensional structure [27]. Pauling et al. proposed prior that there have to be regular, stabilizing elements (such as α-helices), though when protein structures were elucidated, the arrangement of these stabilizing elements were found to be surprisingly irregular [68, 69]. Several questions have been put forward which remain to be answered. Alternative folding pathways have been described for homologous proteins [70]. It is an open question if a general folding pattern can be derived which is relevant for all proteins [71]. Also, there is dispute about which aspects of protein folding are stochastic and which are deterministic [18, 72]. How is the hydrophobic core of a protein established [55, 27]? There are cases where the hydrophobic collapse to a molten globule precedes the formation of secondary structure elements [56]. Several

folding theories have been published which partially preclude or complement one another. However, no consensus has been achieved [73, 74, 12]. Most theories are directly linked to experimental data obtained for specific protein structures. If there is no general folding mechanism, the presented theories may be equally valid, though they may be applicable only to specific proteins.

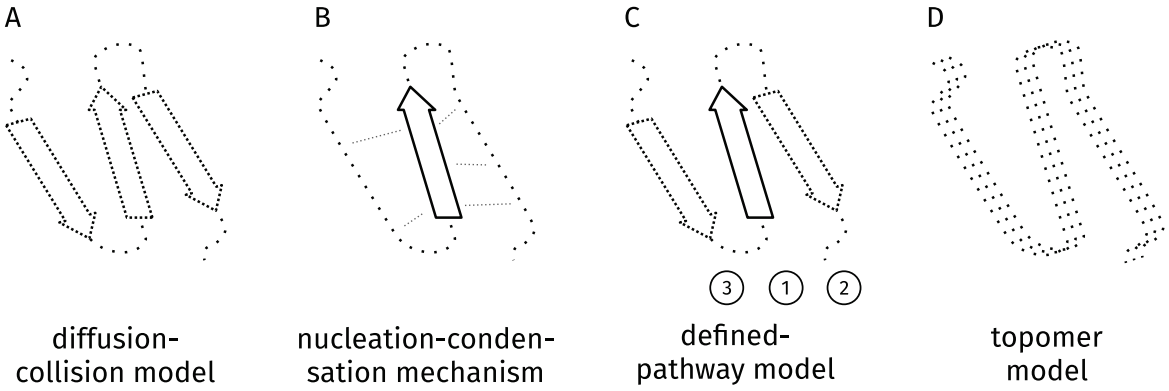

Figure 2.4.: Comparison of Folding Theories

Graphical depiction of folding theories. (**A**) Incipient microdomains can be formed quickly due to their limited size and will diffuse and collide into the native conformation [75]. (**B**) A folding nucleus is established which forms transient contacts to other protein parts which are still unstructured [76]. (**C**) A deterministic, stepwise assembly of autonomous folding units guides folding [19]. (**D**) Within a set of near-native structures the final conformation is searched [77].

Diffusion-Collision Model The diffusion-collision model (Figure 2.4A) proposes that proteins are composed of microdomains for which the correct (partial) conformation can be searched exhaustively [78]. These microdomains are not stable (because of their small size [75]), but rather will diffuse and collide, assembling into the stable, native conformation [78]. Incipient secondary structure elements have been identified as microdomains in myoglobin. They need to be partially established before their relative assembly can begin which is ultimately stabilized by hydrophobic interactions [79]. In other cases, the mere correct orientation of chain fragments or the assembly of hydrophobic clusters may suffice. This may relate to a hydrophobic collapse and a subsequent molten globule state. In agreement with other theories, folding does not depend on individual residues but is rather the consequence of the properties of microdomains (i.e. their sequence composition) and interactions between them [75]. Especially for helix bundles, consolidation of the structure was found to follow a hierarchic variant of the diffusion-collision model [80, 27].

Nucleation-Condensation Mechanism The nucleation-condensation mechanism is based on the (partial) formation of one particular folding nucleus (such as an α-helix) during the transition state of folding (Figure 2.4B). This intermediate is further stabilized by tertiary contacts of the folding nucleus with other parts of the protein which still lack a defined structure [76, 34]. Subsequently, the rest of the protein collapses around this folding nucleus and condensates into a compact conformation which constitutes the native conformation. It is argued that the unstable nature of the folding intermediate prohibits their accumulation thereof during folding. Thus, a fast, two-state folding is achieved. Evolutionary pressure acts on the folding nucleus in the sense that this transient nature has to be preserved [76]. Only a fraction of the protein initiates folding, subsequently consolidation of the folding nucleus and its extension happen concurrently. Secondary structure elements and tertiary contacts are also established simultaneously [34]. A hierarchic interpretation

of the nucleation-condensation mechanism suggests a step-wise assembly of folding intermediates, primarily realized by secondary structure elements [81].

Defined-Pathway Model The defined-pathway model (Figure 2.4C) is motivated by findings of the experimental investigation of proteins (Figure 2.5). Therein, so-called foldons have been identified as autonomous folding units [18, 12, 19, 82]. These fragments encompass 15–35 residues [19] and fold autonomously – no other region of the protein directly supports or hinders their formation [18, 19]. It is encoded in their sequence which parts of the protein initiate the formation of local, ordered structures, e.g. secondary structure elements [83, 84, 13]. These regions decrease in free energy as well as entropy and stabilize the protein during the folding process [85, 13]. This also supports the observation that proteins fold cotranslationally as they are being synthesized by a ribosome and stabilizing tertiary contacts cannot be formed yet [86]. These local structures assemble into the global structure [75, 82, 18]. Tertiary contacts are especially important for the stability of the hydrophobic core of the native structure [87]. The defined-pathway model suggests that the irreducible entities relevant for protein folding are foldons and not individual residues [12, 33].

The energy landscape theory proposes that protein folding is a random process which finds the native conformation of a protein by exploring an energy funnel and gradually decreasing the energy of the current conformation. This exploration is random (though guided by the decrease in free energy and an accumulation of native contacts [85, 88, 89]), but many ways will lead to a correctly folded protein. The defined-pathway model contradicts this view as experimental data indicates the opposite [12, 90]: proteins fold reproducible in the sense that foldons will be established at distinct points in time and this formation is the same if the experiment is repeated. If this assumption is true, then the initial search of the foldon with the highest propensity to fold should be a random process and happen slowly. After that, the formation and relative assembly of further foldons may be guided by the already established structural fragments and will happen faster. The reason is that more information on the correct assembly of the structure is present and the number of conformations to explore is reduced due to the presence of folding intermediates. This leads to the observed two-state folding for most small proteins for which no folding intermediates accumulate [12]. Interestingly, molecular dynamic simulations of protein folding come to a similar conclusion [91, 19, 92].

Topomer Model Topomers (Figure 2.4D) are sets of near-native conformations which do not disrupt the covalent bonds of the peptide backbone. These conformations have been shown to bury surface area in a manner comparable to the native structure which implies that the hydrophobic collapse precedes the sampling for the topomer which resembles the native structure [77]. Analogously to the defined-pathway model, the diffuse search for the topomer ensemble is slow whereas the refinement of a structure with overall correct topology happens fast. The duration of the first time step scales with the number of residues [93].

2.3. PULSE-LABELED HYDROGEN-DEUTERIUM EXCHANGE

The conclusion of the previous section is that the protein folding process is still little understood which is the consequence of a lack of standardized experimental data describing folding intermediates [9, 10, 11], especially difficulties in observing the transition state obscures important characteristics [94, 95]. Various experimental techniques have been established to investigate the folding process (and the transition state in particular) with temporal and spatial resolution. The most promising technique is pulse-labeled hydrogen-deuterium exchange (HDX). Therein, it is exploited that backbone amide groups of amino

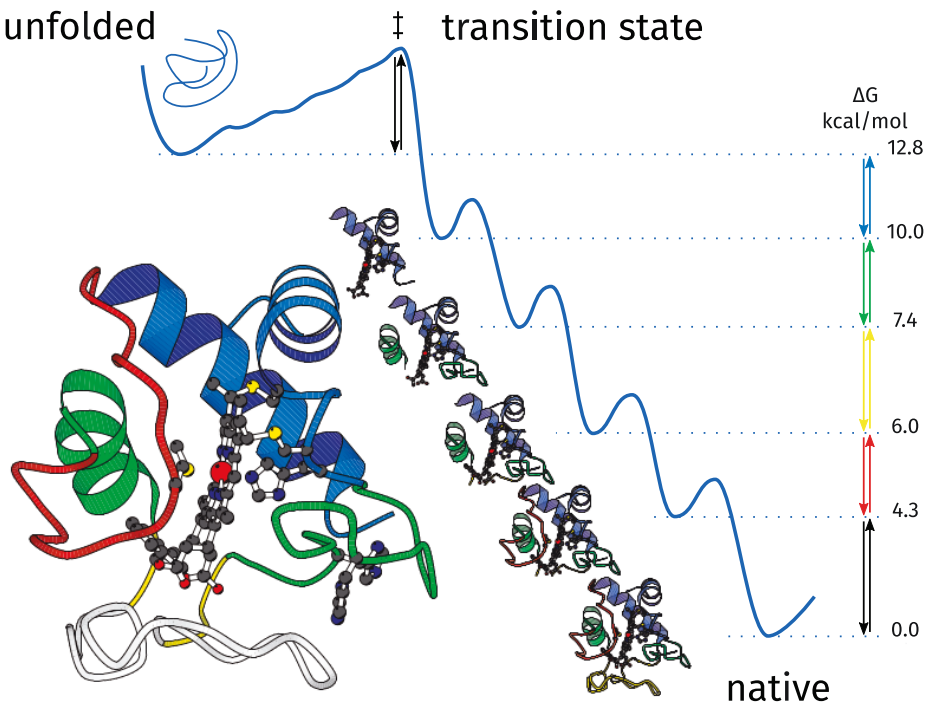

Figure 2.5.: Defined-Pathway Model of Protein Folding

During protein folding an unfolded protein chain passed an energetic barrier – called transition state – and folds into the native conformation. The final structure exhibits low free energy and low entropy. The defined-pathway model [19] is supported by experimental data which implies that protein folding is a deterministic, step-wise process. After the transition state is passed, local stable structures are established and form energetically favorable assemblies (first folding intermediate depicted in blue). This results in a partial decrease in energy. Residues of the blue fragment are considered to fold early in the process (early folding residues). Subsequently, other parts (depicted in green) will fold and are associated to the already folded parts of the structure. The order of this process is determined. The folding intermediates are assumed to guide and stabilize the process. Figure adapted from [19].

acids can exchange their hydrogen atom for deuterium atoms. This exchange depends on the accessibility of the amide group: residues in the core of the protein do not interact with the solvent and are therefore not susceptible to this exchange. The same is true for residues in stable secondary structure elements such as α-helices. In contrast, residues at the surface of the protein and those in unstructured, disordered regions will be readily accessible by the exchange. By combination with nuclear magnetic resonance (NMR) or mass spectrometry (MS) experiments, it can be assessed whether an exchange occurred at a particular residue [87, 96, 12, 19]. HDX provides information on the local and tertiary structure, structural change or dynamics, as well as energetics [12]. An additional level of information can be added when proteins are denatured in a controlled manner. Denaturation agents such as GuHCl disrupt the formation of hydrogen bonds in the solvent and act as chaotropic substances: unfolded states can be introduced to artificially initiate protein folding and unfolding events [97, 98].

Factors which influence the rate of HDX are diverse [12]: the process depends on pH, temperature, surrounding residues, isotope effects, and ionic strength [99, 100, 101]. Li and Woodward pioneered a protocol to compare HDX experiments despite changing experimental conditions or techniques [87]. Their methodology was later employed for the design of the Start2Fold database [8] which currently encompasses HDX data of 57 proteins in a standardized manner. The database covers all structural protein families present

in class, architecture, topology, homologous superfamily (CATH) and structural classification of proteins (SCOP) [11]. However, the size of the deposited proteins [13, 11] varies from 56 to 394 residues which likely makes this resources only relevant for the folding of similarly small proteins. However, it is quite common to simplify the protein folding problem to single-domain structure of this size [32]. The knowledge of folding characteristics is difficult to interpret because to truly understand such a specific aspect of folding, the process itself has to be understood fully [12].

2.3.1. EARLY FOLDING RESIDUES

Starting from a denatured protein, folding conditions are gradually established until the protein refolded completely. The resulting folding trajectory can be studied by HDX. Residues become protected when their amide group is isolated from the solvent as the effect of other residues surrounding them. During protein folding, the spatial neighborhood of a residue may be altered. Thereby, especially the formation of hydrogen bonds involving the amide group of the backbone is relevant. Residues which are protected from the exchange at the earliest stages [96, 20, 18, 13] are called early folding residues (EFR). Residues which are protected only at later stages or not at all are referred to as late folding residues (LFR). The protection of amide groups occurs at an exceedingly fast timescale below 1 s which are difficult to track experimentally. In some cases, the experimental signal of EFR may not be the effect of the formation of hydrogen bonds but rather be the mere result of undirected physical chemistry. In other cases, buried groups are still accessible to the exchange [87]. Several models try to explain this observation; but in general breakage of hydrogen bonds is required to allow exchange of the hydrogen atom of the backbone amide group [12]. Proteins are flexible structures: maybe multiple, non-cooperative fluctuations allow solvent molecules to access buried amide groups. Another assumption is that whole fragments of a protein will unfold from time to time. Fragments may also adopt different conformations as the free energy of the optimal state is only marginally favorable. Despite all uncertainty, the relevance of the experimental signal can be assumed to be more reliable when several residues in spatial proximity exhibit similar tendencies [87].

EFR were shown to initiate the folding process and the formation of secondary structure elements [13] or even larger autonomous folding units [18]. They were found to form tertiary residue contacts early during the folding process [87]. EFR tend to be conserved, non-functional residues [22]. In contrast, LFR may be relevant during later stages of the folding process, implement protein function, or be mere spacers between protein regions. In a previous study [20], EFR have been shown to exhibit lower disorder scores and higher backbone rigidity. Regions with high backbone rigidity are likely to constitute ordered secondary structure elements [20, 13] and this tendency is manifested in local sequence fragments [83, 84]. Especially aromatic and hydrophobic amino acids were linked to ordered regions of proteins [20]. Subsequently, it was shown that EFR are likely buried according to their relative accessible surface area (RASA) [13]. EFR have been shown to feature the largest number of residue contacts in the folded structure and were linked to an increased evolutionary co-variation [11].

The annotation of EFR exists only for a small number of proteins. EFoldMine [11] is a classifier that predicts EFR from sequence. Due to the nature of the trained models [20, 11], it is still unclear what characteristics cause EFR to fold first [13]. Furthermore, early folding events are enigmatic [102, 103, 72]. EFR are a resource to address this question: are the experimental signals of EFR transient? This would imply that EFR are only relevant in the early stages of the folding process and will not exhibit distinct characteristics in the successfully folded, native conformation. Ways to determine the most relevant structural features for EFR are presented in Chapter 5 and Chapter 6.

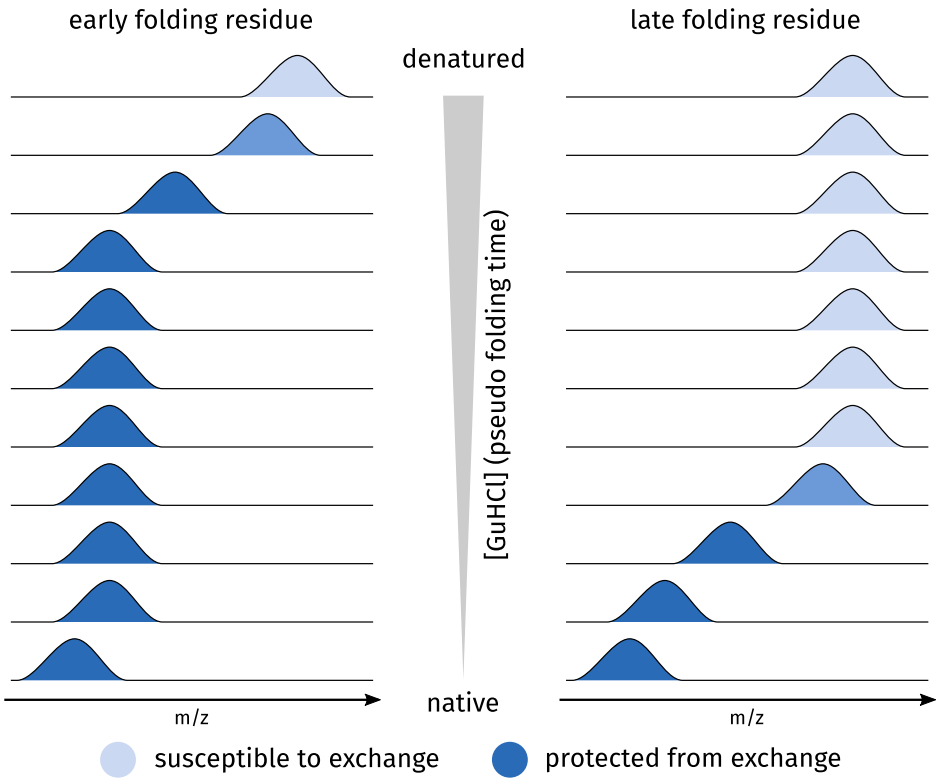

Figure 2.6.: Experimental Identification of Early Folding Residues

Pulse-labeled HDX can be used to identify residues which fold first in the protein folding process. The hydrogen atom of the backbone amide group of an amino acid is exchanged with deuterium or vice versa. Whether this exchange happened or not can be measured by techniques such as MS or NMR and related to one particular amino acid. The temporal resolution of the folding process is achieved by denaturation agents which allow minute control over the folding state of a protein. By combining various states, the folding process can be observed with a pseudo time attached to each observation (y-axis). In the initially denatured protein chain all residues are susceptible to HDX (light blue). So-called EFR are residues which become protected from this exchange first. Protection (dark blue) is the result of residues being shielded by surrounding residues – their amide group is no longer accessible by the solvent and no exchange can be observed because backbone hydrogen bonds have been formed. In contrast, there are LFR which become protected late or not at all [87].

2.3.2. HIGHLY STABLE RESIDUES

In addition to EFR, highly stable residues (HSR) were identified in HDX experiments. This particular type of residue is observed by denaturating protein structures by chaotropic agents such as GuHCl [97]. The higher the concentration of the denaturant, the more the formation of hydrogen bonds in the protein structure is disrupted. This leads to an increasing tendency of the protein to lose its structure and adopt the conformation of an extended chain (Figure 2.7). Most residues will lose their structural integrity quickly: they are referred to as unstable residues (UR). In contrast, certain sequence fragments form highly stable secondary structure elements which do not necessarily fold early but are exceedingly stable. Residues therein are HSR as they resist unfolding pressure. They are the last residues in a protein which will give way to a random structure.

The relevance of HSR is difficult to assess. It has been stated that protein unfolding events are way easier to understand than folding events. Especially identifying an unfolding pathway is more feasible than finding a folding pathway. Because of the symmetry of the process, this investigation in the opposite direction may provide new or complementary

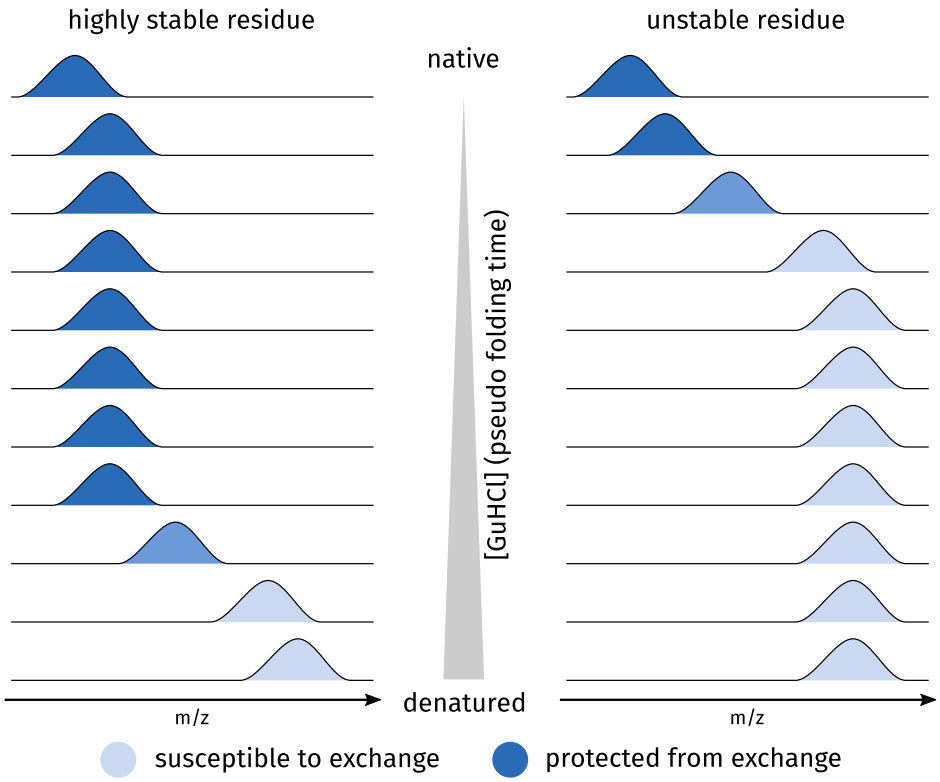

Figure 2.7.: Experimental Identification of Highly Stable Residues

Pulse-labeled HDX can also be used to identify residues which are highly stable with respect to unfolding events. Starting from the native structure, the concentration of denaturation agents is increased so that the protein becomes increasingly unstable. Some regions of the protein will almost instantly lose their structural integrity (UR) and become susceptible to the exchange of the hydrogen atom of their backbone amide group. In contrast, HSR resist unfolding even at high concentrations of a denaturation agent as indicated by their protection from the exchange.

insights [32]. The role of HSR for the structural integrity in proteins is presented in Chapter 4.

2.4. OTHER EXPERIMENTAL TECHNIQUES TO STUDY PROTEIN FOLDING & UNFOLDING

Several other techniques approach the problem from a similar angle and try to identify key residues for the protein folding process or protein stability. Some approaches indirectly accumulate information on the transition state which is not directly observable, while others disturb the native conformation of a protein by pH, heat, or denaturation agents such as GuHCl or urea [104, 105].

ϕ-Value Analysis ϕ-value analysis is based on mutagenesis studies [106]. Residue positions are mutated (e.g. to an alanine amino acid) and the change in energy of the denatured, transition, and native state of the protein is tracked. The derived value states how drastic the introduced mutation changes the folding characteristics of the transition state and can be used to characterize the surroundings of a residue in the context of protein folding. Scores close to 0 indicate no effect on the transition state as well as native conformation. In contrast, values close to 1 represent destabilizing effects. Furthermore, the relation of the free energy of the transition and native state captures whether the studied residue is embedded in an ordered, already folded surrounding during the transition state [106, 107].

Tryptophan Scanning Another way to study conformational change during protein folding is to exploit the fluorescence of the aromatic amino acids phenylalanine, tyrosine, and tryptophan. Of these amino acids, tryptophan is the most relevant candidate for fluorescence studies because it allows for the most precise measurements (e.g. due to its high extinction coefficient). Tryptophan fluorescence is especially suited to characterize folding intermediates because the emission energy of tryptophan changes depending on its embedding in a protein structure. The emission is blue-shifted for residues in the hydrophobic core of the protein, whereas a red-shift is expected for partially buried or exposed tryptophans [108]. Basically, this technique allows to assess whether tryptophan residues show similar characteristics as in the native or denatured state or if folding intermediates exhibit completely different conformations [109].

Circular Dichroism Circular dichroism [104, 110, 111] characterizes the presence of secondary structure elements in protein structures. The relative content of α-helices and β-sheets can be determined especially reliably [104]. This approach exploits the difference in adsorption of differently polarized light by enantiomeric molecules (such as proteins). Different classes of secondary structure elements have distinct spectra and can thus be distinguished. Furthermore, their relative content can be determined, especially when conditions of a protein are varied [110].

This section can be concluded by the statement that no single technique can optimally characterize folding intermediates, but rather an integrative view via the combination of different approaches leads to a deeper understanding of the folding process [108]. This is however hindered by many of the presented measurements being not directly comparable [87]. Also, all techniques provide only a qualitative indication for certain residues to be part of the folding nucleus [87]. Experimental HDX data was chosen to describe folding characteristics of proteins, because the most comprehensive collection of such data is provided by the Start2Fold database [8].

2.5. *IN SILICO* PROTEIN FOLDING

Molecular dynamic simulations can be used to recreate and study protein folding *in silico* [91, 19, 92]. Interestingly, simple lattice models which only consider the hydrophobicity of amino acids in a binary manner yield good results as well [57, 58, 21]. In consequence, it has been argued that protein folding is mainly driven by local interactions which form secondary structure elements which then assemble to form the global structure. Contacts between these secondary structure elements are rather unspecific and are more likely to refine a near-native structure, while not dominating its formation. The structure of proteins is the consequence of them being heteropolymers (i.e. they consist of hydrophobic as well as hydrophilic monomers). This dichotomy and the restriction imposed by the covalent bonds lead to one compact native structure [21]. The concepts leading to successful *in silico* protein folding are potentially also relevant in *in-vivo* folding: even if the *in-vivo* process cannot be observed unconditionally, the problem may be approached by simulating simplified models *in silico*.

3. CONTACT-BASED STRUCTURE PREDICTION

A given protein structure with n atoms can be represented by $3n$ numbers which describe the x, y, and z coordinates of all atoms. This representation can be transformed to a distance map (DM) which captures the Euclidean distance between all atom pairs. This map is a symmetric matrix and contains $n \cdot (n-1)/2$ entries for all possible pairings of atoms. Both representations can be transformed into the other. The main advantage of a DM is that they are invariant to rotation and translation. Also, they can be sparse, i.e. the majority of elements may be missing, while still containing all information necessary to derive the three-dimensional structure [112].

3.1. CONTACT MAPS

A DM can be further simplified by transforming this matrix of real values to a binary matrix which contains two states. This representation is called contact map (CM). It is evaluated whether a distance threshold T (e.g. 8 Å which equals 0.8 nm) is fulfilled for a given pair of atoms i and j:

$$CM[i,j] = \begin{cases} 1, & \text{if } DM[i,j] \leq T \\ 0, & \text{if } DM[i,j] > T \end{cases} \tag{3.1}$$

This leads to a drastic reduction of contained information. Another common simplification is to merely evaluate pairs of amino acids. Residues are represented by their C_α, C_β, or centroid and the distance threshold T is evaluated between these representatives. The coordinates of a protein can be reconstructed from this further reduced representation as well by several computational approaches [113, 114, 16, 115].

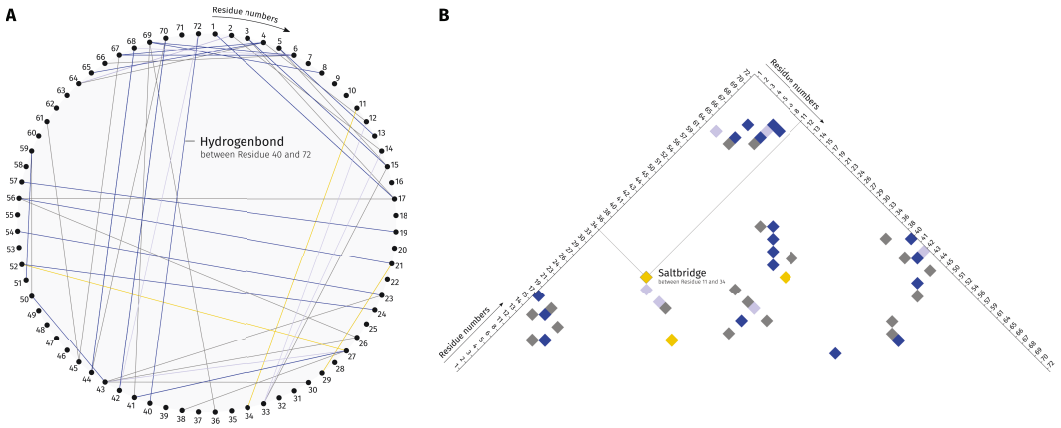

Figure 3.1.: Visualizations of Contact Maps

CMs represent contacts between residues in a protein structure. These renderings are a personal communication with Christoph Leberecht, Florian Kaiser, Sebastian Salentin, Michael Schroeder, and Dirk Labudde. (**A**) The "yarn plot" is a circular representation of contacts between residues. Edges between residues are colored by interaction type as annotated by PLIP [37]. Hydrogen bonds are colored dark blue, water bridges light blue, salt bridges yellow, and hydrophobic interactions grey. (**B**) The "pyramid plot" is a reinterpretation of CMs, which omits redundant information present in conventional, symmetric CMs. Positions which do not feature any contacts to other residues are ignored for visual clarity.

CMs are a welcome simplification of three-dimensional protein structures. Structures are not readily conceived by humans as well as compared to other structures because they are usually placed arbitrarily in the three-dimensional space. CMs capture a high amount of structural information while being inherently invariant to rotation and translation. This makes interpretation and comparison more feasible [112]. E.g., α-helices constitute thick bands around the main diagonal because they are composed of contacts between a residue and

its four successors. In contrast, parallel and anti-parallel β-sheets manifest as thin bands parallel or anti-parallel to the main diagonal, respectively [113]. The binary character of CMs has implications for the prediction of contacts by machine learning techniques since a binary variable is easier to learn than a continuous one [112]. Studies [14, 15, 116, 17] have demonstrated that CMs are tolerant to the deletion of individual contacts. This implies that even sparse representations of a CM can be used for the successful reconstruction of a protein structure since most information in a CM is redundant [14, 15, 17]. This again benefits contact prediction methods as the number of true positive predictions does not need to approach 100%.

Of course, the employed simplification also comes at a price. Reconstruction routines are tolerant to the deletion of contacts and some authors [112] state that the same is true for the introduction of some false positive contacts. Such contacts are not prepresent in the CM of the native structure, but rather errors of contact prediction methods. However, false positive predictions are usually systematic errors and tend to be intrinsically correlated. Also, contacts influence other contacts in a non-local manner. This leads to complex interactions between false positive contacts, which cannot be compensated [112]. Other authors [17] assessed false positive predictions to be in general far more detrimental to the reconstruction fidelity. In order to address this issue quality assessment methods for CM were developed [117]. Another conclusion drawn is that contact correctness is more important than the number of predicted contacts [17]. Dedicated visualization tools were established [118, 119] because of the success of contact prediction methods and the increasing relevance of CMs.

A crucial aspect for the creation of a CM is the used contact definition. What defines a contact between two residues? How are residues represented? In pivotal studies [120, 16], it has been demonstrated that there is no universal contact definition which is optimally suited for all problems. The non-covalent interaction types [37] which stabilize protein structures are highly diverse. Especially hydrogen bonds and hydrophobic interactions play a key role and have been shown to exhibit distinct distance preferences [120, 121]. In other cases [122], structures such as PDB:1xcp_B cannot be successfully reconstructed with the standard contact definitions but require a roughly doubled distance threshold of 16 Å. Reconstruction fidelity also increases when additional information such as the distribution of secondary structure elements is considered. The same is true for the consideration of non-contacts, i.e. pairs of residues which do not share a contact in the CM [123].

CMs have a wide range of applications for structural biology [124]: especially for the identification of domain boundaries, as basis for molecular replacement routines, for the interpretation of X-ray crystallography (x-ray) and NMR data, the identification of functional sites, and the assessment of crystal packing. Other studies describe the utilization of CMs for the analysis of protein dynamics [125]. They are also used to speed up molecular dynamic simulations [126] as well as fragment-based structure prediction routines [127]. Both approaches try to recreate the protein folding process *in silico*. Dedicated studies approach the protein folding problem by CMs [128]. CMs have also been shown to be valuable tools for the comparison of structures [129] and the description of protein-protein interfaces [130]. In [112], studies on the small-world feature of proteins are reviewed. CM can be represented as residues graphs and therein the number of edges is relatively small while the overall connectivity of the graph is still high. This characteristic is the result of the small-world feature wherein a small number of vertices are crucial hubs within the graph. In summary, CMs are one of the most promising techniques in structural biology and the rapid grow of the surrounding ecosystem captures this.

3.2. CONSTRAINT-BASED *AB INITIO* STRUCTURE PREDICTION

The all-atom reconstruction of a protein structure from a sparse CM is a difficult problem. Initially, structures were reconstructed by applying the Lagrange theorem [131] or stochastic optimization [128]. The latter approach evaluates an objective function which captures how well the set of given constraints is fulfilled and optimizes this function to find a likely conformation. An essential characteristic of the reconstruction strategy is how well it can deal with errors in a CM as present in real world applications [112]. Therefore, several reconstruction algorithms [132, 133, 134, 116] have been proposed which provide an increased performance, applicability, and error tolerance.

A state-of-the-art algorithm for the reconstruction of the tertiary structure from a CM is CONFOLD [114]. The algorithm has native support to incorporate defined secondary structure elements by evaluating distance restraints, dihedral angles, and hydrogen bonds. Conventional algorithms only represent secondary structure elements by distance constraints. Especially β-sheets can be modeled with higher accuracy. Another crucial consideration is the number of enforced distance constraints. The actual reconstruction routine is an adaption of the distance geometry simulated annealing protocol of CNS [132]. Specifically, more minimization steps are performed and the number of backbone atoms for which constraints are considered was increased. Also the protocol was adapted to support weighted constraints and additional information provided by contact prediction techniques which state the reliability of predicted contacts. The gathered information is used to create 20 models which are then scored by the default routines provided by CNS. Unsatisfied contacts are filtered, and a second stage reconstruction may be performed which again yields 20 models. By default, the five top-scoring models are provided as output of the algorithm [114].

An established strategy to score the fidelity of reconstructs is the superimposition with the native structure used to compose the initial CM [14, 15, 17, 135], e.g. by TM-align [136]. The quality of the reconstruct can be quantified by scores such as the root-mean-square deviation (RMSD) or template modeling score (TM-score) [137, 138]. In this thesis, RMSD values below 4 Å are regarded to resemble the native structure and thus constitute a successful reconstruction. Analogously, a TM-score above 0.5 represents the correct fold [138].

3.3. REVOLUTION BY COEVOLUTION

Coevolution techniques revolutionize how scientists approach structural bioinformatics [139]. One such technique is the direct coupling analysis (DCA) [140, 141] which was primarily designed for the protein structure prediction. Subsequently, it has been applied for the prediction of membrane protein structures [142], protein complexes [143], and mutation effects [144].

The DCA makes use of the fact that a large amount of genomic information has been collected. Oftentimes there is a evolutionary connection between sequences as it is the case for protein families. This imposes constraints on the sequences because they cannot change freely but rather have to conserve their function and structure. This results in similar sequences within protein families which can be aligned in a multiple sequence alignment (MSA). Positions within this MSA may change seemingly arbitrarily while others are conserved and cannot change without compromising protein function. Ultimately, the one constraint on proteins is that they must be biologically functional. Otherwise, an organism would not be viable, or the useless gene would not be expressed anymore. This leads to the conservation of crucial sequence positions – over the course of evolution, these positions remained unchanged. Commonly, they can be associated to key positions regarding protein function. A more interesting case is when defined residue positions change in dependence of a second position far away at sequence level (Figure 3.2). Therein, a mutation

at one position would likely result in a dysfunctional protein; however, the detrimental effect can be compensated when the corresponding position changes accordingly. Such coevolution effects can be observed in MSAs and are the basis of coevolution techniques. It is remarkable that positions far apart at sequence level exhibit such complex covariation. The drawn conclusion is that the coupling of both residues is the direct consequence of them being spatially close. Mutations to one positions need to be compensated by the coupled position because otherwise their contact in protein structure would be lost. Ever since this argumentation was established, it was refined for the application in structure prediction routines. Most importantly, this approach does not require knowledge of template structures [140, 141] such it is the case in homology modeling [145].

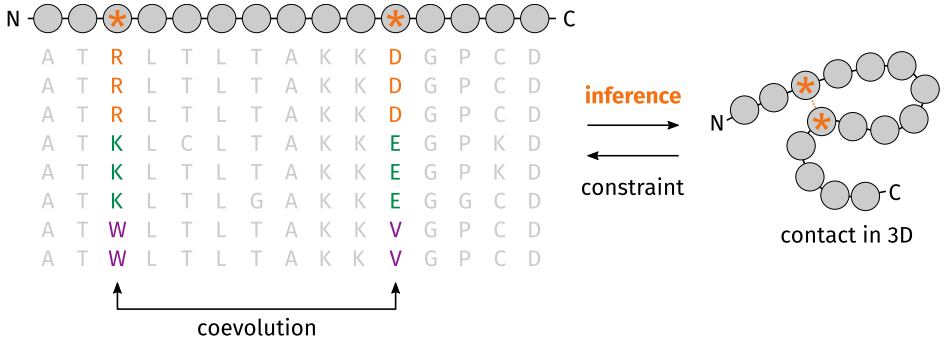

Figure 3.2.: Contact Prediction by Coevolution

Coevolution techniques [140] were designed for protein structure prediction and make use of the fact that many protein sequences are known. By searching for homologous sequences and aligning them, the conservation of individual residues of the query sequence can be assessed. Some positions are conserved, others change freely. A small number of defined sequence positions changes in dependence of the other (one-letter-codes rendered in color). These positions are coupled: change at one position has to be compensated by an according change at the other sequence position. This implies spatial neighborhood of these coupled residues. Subsequently, these coupled positions can be used as constraints of structure reconstruction algorithms. Figure adapted from [140].

Despite the simplicity of the DCA, the concrete realization is disputed. How are coevolving positions represented? The initial approach was based on the local mutual information (MI) measure. Let there be a residue pair of positions i and j. Their MI is defined as:

$$MI_{ij} = \sum_{A_i,A_j=1}^{q} f_{ij}(A_i, A_j) ln \left(\frac{f_{ij}(A_i, A_j)}{f_i(A_i) f_j(A_j)} \right) \tag{3.2}$$

Therein, MI_{ij} is a difference entropy which evaluates the co-occurrence frequency $f_{ij}(A_i, A_j)$ of the amino acids A_i and A_j observed in the alignment at positions i and j and the distribution $f_i(A_i)f_j(A_j)$, when no coupling is present. Since then, it has been shown that the measure based on MI is not optimal as it has a local character (i.e. MI_{ij} is independent of all other possible pairs). Due to this locality, the model is highly sensitive to transitive couplings. If there is covariation between residues A and B as well as between B and C, then any local model will also report covariation between residues A and C. This implies spatial proximity of both residues, even though the observed covariation is the mere consequence of an inadequate measure [140]. To address this issue, a generalization of the Ising model was established. The conventional Ising model describes ferromagnetism and considers the spin of each particle which adopts two distinct states. To quantify covariation in MSAs, 21 states are considered (for the canonical amino acids and gaps in the alignment). The major difference of the direct information (DI) and the MI measure is that the local frequency count $f_{ij}(A_i, A_j)$ is replaced by the twofold constrained pair probabilities $P_{ij}^{Dir}(A_i, A_j)$:

$$DI_{ij} = \sum_{A_i,A_j=1}^{q} P_{ij}^{Dir}(A_i, A_j) ln \left(\frac{P_{ij}^{Dir}(A_i, A_j)}{f_i(A_i) f_j(A_j)} \right) \tag{3.3}$$

P_{ij}^{Dir} can be obtained by evaluating residue couplings and the individual amino acid frequencies. The result is a model which is consistent for all pairs of i and j. In [140] it has been demonstrated that this model is far more sensitive to true positive contacts and also captures more tertiary contacts which have been proven to be crucial for protein folding [146, 147] as well as structure prediction [15, 140]. Transitive contacts are still difficult to detect [148]. In other cases, sequence conservation may be caused by protein function rather than protein structure [149]. EVfold [140, 141] was chosen to represent coevolution techniques in this section due to its pivotal role for the field. Other implementations are e.g. PSICOV [150], CCMpred [151], Gremlin [152], or plmDCA [153]. The major limitation of coevolution techniques is that they require a good MSA based on many reasonably diverse sequences. If no or too few homologous sequences can be found, residue contacts cannot be predicted [135]. Coevolution techniques also have difficulties with predicting membrane proteins correctly [135, 124, 123]. To address this issue, dedicated methods had to be designed [142].

3.4. CONTACT PREDICTION BY SUPERVISED MACHINE LEARNING

Another way to predict residue contacts from sequence are supervised machine learning (SML) techniques such as deep learning. The public was stunned when Google's AlphaGo beat humans in one of the most difficult games there is [135]. In conventional statistics, correlations between carefully defined variables are assessed to gain a bottom-up understanding of a process. Machine learning uses these variables to directly predict the outcome (e.g. a class label) of the process. A detailed understanding of why a certain prediction is made may not be provided. Deep learning makes things even more abstract: the variables used for training do not have to be highly descriptive for the process to be modeled. Instead, deep neural networks can work with simple input variables and will use this provided information to compose complex features which are then used for the classification task. In terms of protein structure prediction, there is homology modeling which explicitly searches for template structures which are subsequently used to create models of the query protein sequence [145]. Machine learning techniques use features derived from sequences such as sequence profiles, predicted secondary structure elements, and coevolutionary information [154]. These features are used for the training process. Deep learning techniques are also based on sequence features but employ convolutional operations to internally derive features which provide a better basis for the learning process.

One remarkable deep learning approach for the prediction of CMs is RaptorX [135]. Therein, the CM of a native protein structure used for training is interpreted as an image and allows to employ knowledge of the computer vision community: each pixel is either active or inactive and captures whether there is a contact between the corresponding residues. The implementation consists of two residual neural networks (Figure 3.3). When the number of hidden layers increases, networks become more complex, and training becomes more difficult as well. To minimize the error function, all weights in the neural network must be adapted appropriately [135]. This is done by quantifying the error of the output layer and propagating it back to the input layer. The longer the distance between input and output layer is (i.e. the deeper the network), the more difficult does this process of back-propagation become. Because the connection of input and output layer is so complex, the overall architecture is prone to degeneration. Residual neural networks address this problem of deep architectures by introducing an additional direct route between input and output layer (the

arrows passing the 1d and 2d conv layers in Figure 3.3). This prevents degeneration of deep networks since it provides knowledge of the initial input to all layers regardless of actual depth in the architecture. In consequence, layers do not have to find a optimal operation out of the blue but can refine the output of their preceding layer because more context is provided [155].

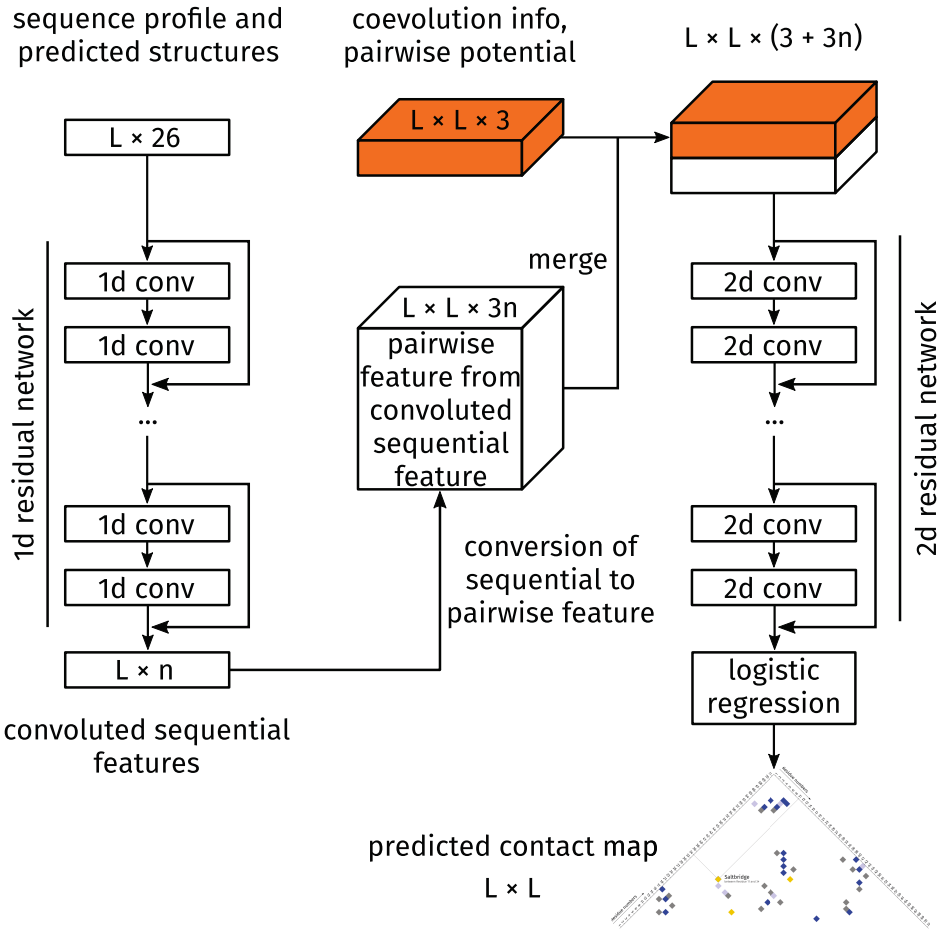

Figure 3.3.: Contact Prediction by Deep Learning

RaptorX [135] (a SML technique) uses sequences features to predict a matching CM for a sequence of length *L*. *n* describes the number of features at a particular step. This concrete architecture is realized by two residual neural networks. Conventionally, a layer must find an appropriate mapping between its input and output without additional information. The residual architecture additionally provides what the preceding layer did and allows the next layer to fine-tune that output rather than starting from scratch [155]. In the one-dimensional part sequence features such as predicted secondary structure elements and accessible surface area are convoluted to create more descriptive features. A matrix is created from this vector by employing the outer product. Additionally, it is enriched with pairwise features between residue positions such as coevolutionary information, a contact potential, and a distance potential. This matrix is again subject to a series of convolutional operations and finally used as input of a logistic regression which predicts the probability of two residues being in contact. Figure adapted from [135].

The input of RaptorX (Figure 3.3) is a vector which describes each residue position by 26 features (i.e. sequence profile, secondary structure elements predicted from sequence, and predicted accessible surface area). By a huge number of residual layers this feature vector is convoluted, and the sequence context of each residue position is learned. The ultra-deep architecture of the network allows to identify relations between residues at sequence level. Subsequently, the outer product of this vector is composed which yields a $n \times n$ matrix. This

matrix is enriched with additional input features in the form of coevolutionary information, a contact potential, and a distance potential. Again, this matrix is fed to a series of residual layers which convolute the initial input. This time, the neural network learns residue co-occurrences and higher order contacts which resemble the two-dimensional context of a residue. The derived features describe the pairing of all residues and a logistic regression is used to transform them to a probability which states how likely it is that these two residues exhibit a spatial contact. The predicted CM can be used to reconstruct an all-atom model of the protein sequence by standard means. This method has been demonstrated to generalize the problem surprisingly well. It was trained solely on globular proteins, but still showed superior performance when employed on membrane proteins. Coevolution techniques such as the DCA learn the sequential context of a residue position within the context of the corresponding protein family. In contrast, this approach evaluates all provided sequence information which potentially includes non-homologous sequences and therefore may be able to identify general patterns in protein sequences. Another improvement to previous methods is that the whole network is trained simultaneously (rather than separating 1d and 2d operations) which allows layers to fix errors made by preceding ones [135].

RaptorX [135] was chosen to represent supervised machine learning techniques in this section because of the diversity of applied machine learning techniques and the superior performance. Other implementations are e.g. SVMSEQ [154], CMAPpro [156], PconsC2 [157], MetaPSICOV [158], PhyCMAP [159], or CoinDCA-NN [160]. Still, RaptorX considers coevolutionary information and it is very important for the accuracy of the contact prediction [135]. In consequence, DCA and SML are not opposed approaches but can rather complement one another. Actually half of the previously mentioned SML techniques [158, 157] directly incorporate evolutionary information and can be considered hybrid methods [124]. The same is true for DCA methods [152], which include SML data.

3.5. THE STRUCTURAL ESSENCE OF CONTACT MAPS

With the emergence of CMs and success stories linked to them, it became evident that a fine-grained interpretation of a CMs is still missing. It is especially unclear why certain contact definitions capture a protein fold successfully and others do not [16]. The most prominent question is what the smallest subset of a CM is, which contains all information to resemble the fold of a protein. It is hypothesized that consideration of a minimal number of contacts as constraints may make structure prediction techniques more efficient [15].

Chen et al. [14] approached this question by defining a dataset, randomly selecting fractions of all contacts, and tracking the influence of this selection on the reconstruction performance. They found that reconstruction fidelity decreases as the number of considered contacts decreases. In their study, 70% of native contacts allowed for successful (i.e. the correct fold is resembled) reconstructs. Another key finding is that random selections of contacts outperform any proposed rational selection of contacts. Good reconstructs were achieved by a combination of local and tertiary contacts. However, no general pattern has been derived [14]. This emphasized the limited understanding for the information captured by CMs.

Subsequently, Sathyapriya and coworkers [15] refined the study by Chen et al. [14]. The coined the term "structural essence" for the minimal set of fold determining contacts. Most prominently, they were able established a rational selection of contacts (called "cone peeling") to reduce the fraction of considered contacts down to 8%. Their contact definition is relaxed compared to that of Chen et al. and they also considered higher RMSD values to be successful reconstructions. The cone peeling selection is based on the observation that many contacts of a map are redundant, i.e. other contacts (or combinations thereof) may contribute the same information for the reconstruction process. Contacts can be eliminated

if they share a common neighborhood or if they are local contacts [15, 126]. Effectively, cone peeling allows the same reconstruction quality with half the contacts. This study also showed that tertiary contacts are more informative (in terms of being the structural essence) than short-range ones, but they are not the exclusively relevant for good reconstructions. In fact, a random selection of contacts does not perform worse than the explicit consideration of the top-scoring tertiary contacts as determined by the cone peeling algorithm. Thus, good reconstructions must encompass both local and tertiary contacts [15].

A general implication of these [14, 15] and related [114] studies is that contacts are not equally important. But no method was established which aims to define the importance of individual contacts in the context of a CM of a native protein structure. And exactly that will become a central aspect of this thesis in Chapter 4.

Part II.

NEW APPROACHES & RESULTS

4. SUPERIMPOSING PROTEIN FOLDING AND STRUCTURE PREDICTION BY CONTACT MAPS AND EXPERIMENTAL FOLDING CHARACTERISTICS

The annotation of EFR and HSR provided by the Start2Fold database [8] constitutes valuable information to understand the protein folding problem and has also implications for the prediction of protein structures. EFR (Figure 4.1A) initiate the formation of stable local structures starting from the denatured protein chain [20, 18]. In contrast, HSR (Figure 4.1B) constitute regions in the native conformation [161] which are resilient to unfolding events (e.g. as natural phenomenon [162] or change in temperature or pH [105]). Both EFR and HSR are key to understand the protein folding process [163, 13]. CMs are the cornerstone of contemporary structure prediction methods. The surrounding ecosystem of reconstruction algorithms may elucidate the protein folding process by pinpointing the most important contacts for structural integrity. Additionally, the relevance of EFR and HSR in the context of protein structure prediction provides qualitative insights. However, the relevance of EFR on the structural integrity of a protein structure is little explored. One reason is that it is currently not possible to assess the role of a contact or residue regarding the structural integrity of a protein; especially an *in silico* approach suitable for large-scale studies is needed to assess the relevance of EFR and HSR.

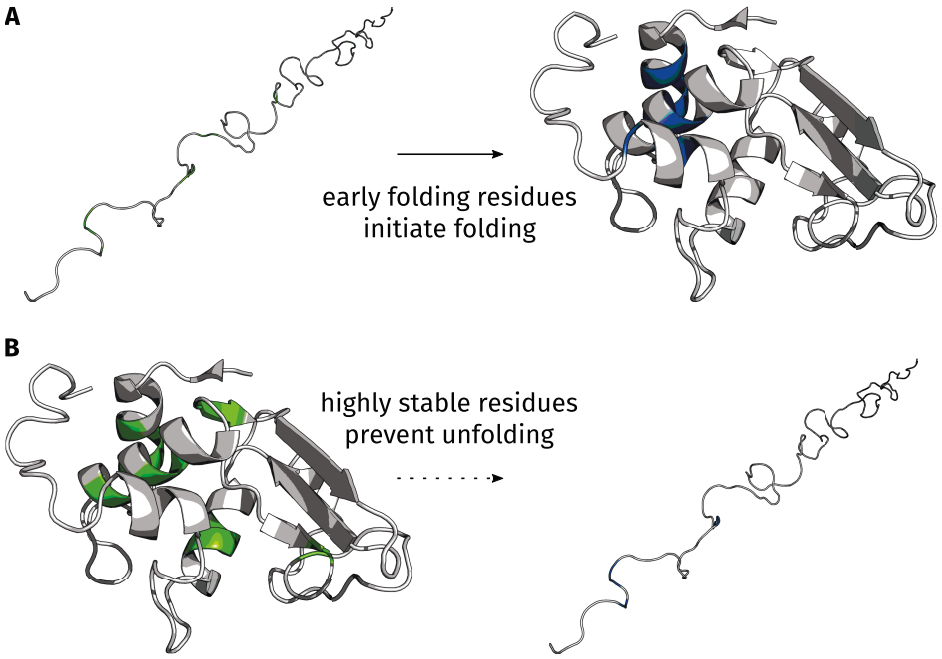

Figure 4.1.: The Relevance of Early Folding and Highly Stable Residues

Most proteins adopt a native conformation autonomously in the process of protein folding [27, 28]. (**A**) A small number of EFR (depicted in blue) initiate the folding process as their surroundings change before that of other residues [13]. (**B**) Analogously, folded proteins can be analyzed with respect to their stability. HSR (depicted in green) occur in regions which are particularly resilient to unfolding events [161].

CMs do not only contain the information needed for protein structure prediction, but they also are potential tools to describe the fundamentals of protein folding. In 2007, Chen et al. [14] pioneered the search for the most relevant contacts of a CM and wanted to determine the minimal set of contacts which captures the fold of a protein. Therefore, they represented proteins by CMs and selected random subsets with varying coverage. These subsets were then used as constraints in a structure reconstruction algorithm, the result was aligned to the native structure, and its fidelity was assessed by the RMSD. As the number of constraints increased (i.e. more contacts of the native CM are considered), the RMSD decreased because the reconstructs resembled the native structure increasingly well. A reconstruction is considered successful when the RMSD to the native structure is below a

certain threshold and likely to resemble the correct fold [17, 15, 14]: in this thesis 4.0 Å are considered as cutoff. Good reconstructions have been shown to depend on a delicate balance of sequentially neighbored and sequentially separated contacts [14]. Sathyapriya et al. [15] extended the study of Chen et al. and coined the term "structural essence" for the minimal set of fold defining contacts. They demonstrated that 8% of all contacts allow for the reconstruction of the correct fold of a protein because most information in a CM is redundant. Furthermore, a rational selection of contacts can outperform a random selection of equally many contacts with respect to reconstruction quality. However, such a configuration is difficult to compose [15]. Duarte et al. showed that consideration of all contacts leads to reconstruction qualities around 2 Å [16].

Several studies identified a small number of key residues for the *in vitro* folding process. Is the same true for *in silico* folding: are some constraints more important than others? Are positions featuring evolutionary couplings crucial for reconstructions? For a long time *in silico* folding simulations improved the understanding of the protein folding process [21, 57], potentially CMs provide an even more tangible connection of both aspects.

4.1. MATERIALS & METHODS

4.1.1. DATASET CREATION

The Start2Fold database [8] provides results of pulse labeling hydrogen-deuterium exchange experiments. Of this dataset, 5,173 contacts of 2,529 residues were evaluated. Positions without native contacts were ignored. The Start2Fold database was chosen because it provides a standardized annotation of EFR which initiate the folding process [13, 11] and HSR which exhibit significant resilience to unfolding events [8]. This dataset encompasses all major CATH and SCOP classes. Thus, the structural relevance score was assessed using a dataset of proteins for which the folding characteristics are fairly well-understood. The size of proteins in the dataset varies from 56–164, which emphasizes relatively small proteins. The covered fold classes are diverse, but present proteins tend to be single domain proteins with fast folding kinetics [11]. Entries without EFR annotation were ignored, even when information on HSR was present. BioJava [164, 165] implementations of the algorithm of Shrake and Rupley [166] and dictionary of protein secondary structure (DSSP) [167] were used for RASA and secondary structure element computation respectively. Residues were considered buried when their RASA was below 0.16 [168]. Evolutionary couplings were computed by the EVfold web server [140, 141].

4.1.2. ANNOTATION OF RESIDUE CONTACTS

A pair of residues was defined to be in contact when the distance between their C_α atoms was less than 8 Å. CMs were created based on this contact definition while ignoring local contacts between residues less than six positions apart at sequence level. The remaining tertiary contacts were considered short (sequence separation of 6–11), medium (12–23), or long (>23) [169]. Non-covalent interactions (hydrogen bonds and hydrophobic interactions) were annotated by PLIP [37].

4.1.3. STRUCTURE RECONSTRUCTION AND PERFORMANCE SCORING

CMs (or subsets thereof) were reconstructed to all-atom models by CONFOLD [114]. Secondary structure information was annotated by DSSP [167] and provided as input of the reconstruction routine. By default, CONFOLD creates 20 reconstructs and selects the five

top-scoring ones as output. The selected reconstructs were then superimposed with the native structure by TMalign [136]. Their dissimilarity was measured by the RMSD.

4.1.4. THE STRUCTUREDISTILLER ALGORITHM

The StructureDistiller algorithm (Figure 4.2) evaluates the structural relevance of individual contacts in the context of a set of other contacts. By selecting 30% of the native contacts of a map, baseline reconstructs can be created which resemble the protein fold and are highly sensitive to the toggling (removal or addition) of an individual contact. The performance of the baseline reconstructs can be quantified by a structural alignment to the native structure. Analogously, the performance can be measured for the toggle reconstructs, which represent the change of one particular contact. By comparing the performance of a toggle reconstruct with its corresponding baseline reconstruct, the structural relevance of all contacts is quantified.

The StructureDistiller algorithm is represented in Algorithm 1. A protein structure S_{native} in PDB format is the input. Structure files should encompass single domains of a single chain. The corresponding CM C_{native} is created. C_{native} constitutes the set of all contacts in the structure which will be evaluated.

Fractions equal to 30% of C_{native} are then randomly selected (Figure 4.2A). The structural relevance of a contact depends on all other contacts used for a reconstruction. No effect can be expected when a contact is considered which contributes no additional, but only redundant information [15]. The creation of random subsets of C_{native} is performed with a redundancy r of 10. The resulting subset of contacts $C_{\text{baseline},i}$ is used to create the baseline reconstructs $S_{\text{baseline},i}$. The average $\text{RMSD}_{\text{baseline},i}$ of each created subset $C_{\text{baseline},i}$ is tracked with respect to S_{native} (Figure 4.2C). These subsets are highly sensitive to the removal and addition of a single contact and the basis for all further computations.

All contacts of C_{native} are now evaluated regarding their structural relevance by pairing each contact to each baseline subset of contacts $C_{\text{baseline},i}$. For each pair, it is determined whether the current contact c is element of $C_{\text{baseline},i}$. If so, c is removed from $C_{\text{baseline},i}$, else c is added to the corresponding subset. The change in reconstruction performance can be quantified by this toggling of a contact (Figure 4.2B): the modified subset $C_{\text{toggle},i}$ is again used for a reconstruction and $\text{RMSD}_{\text{toggle},i}$ is used to describe its quality (Figure 4.2C). The average improvement of the reconstruction with knowledge of the contact c is tracked by ΔRMSD_c. $\text{RMSD}_{\text{baseline},i} - \text{RMSD}_{\text{toggle},i}$ is evaluated when c was removed from the subset, the expression is flipped when c was added. The structural relevance of individual residues is the average of all ΔRMSD_c of contacts this residue participates in. Positive structural relevance scores represent contacts which increase reconstruction fidelity while negative scores occur for contacts hindering reconstruction.

The runtime of StructureDistiller scales with the number of contacts (n_c) in the initially created map C_{native}, the chosen redundancy (r), and a constant factor R required for a particular reconstruction task by the CONFOLD algorithm [114]. This means that runtime scales linearly regarding the number of contacts (n_c). However, the number of contacts in a protein is not linearly related to the number of residues (n_r). In the worst case every residue is connected to every other residue: then $0.5 \cdot n_r \cdot (n_r - 1)$ contacts have to be evaluated. Comparable studies [14, 15] assumed that all possible combinations of contacts have to be evaluated to assess the relevance of individual contacts. This emphasizes the elegance of the proposed algorithm which merely depends on the actual number of contacts.

$$\mathcal{O}(n_c \cdot m \cdot R) \tag{4.1}$$

The individual reconstruction tasks are distributed among worker threads which allows for efficient parallelization. Using a conventional workstation, computation on proteins with

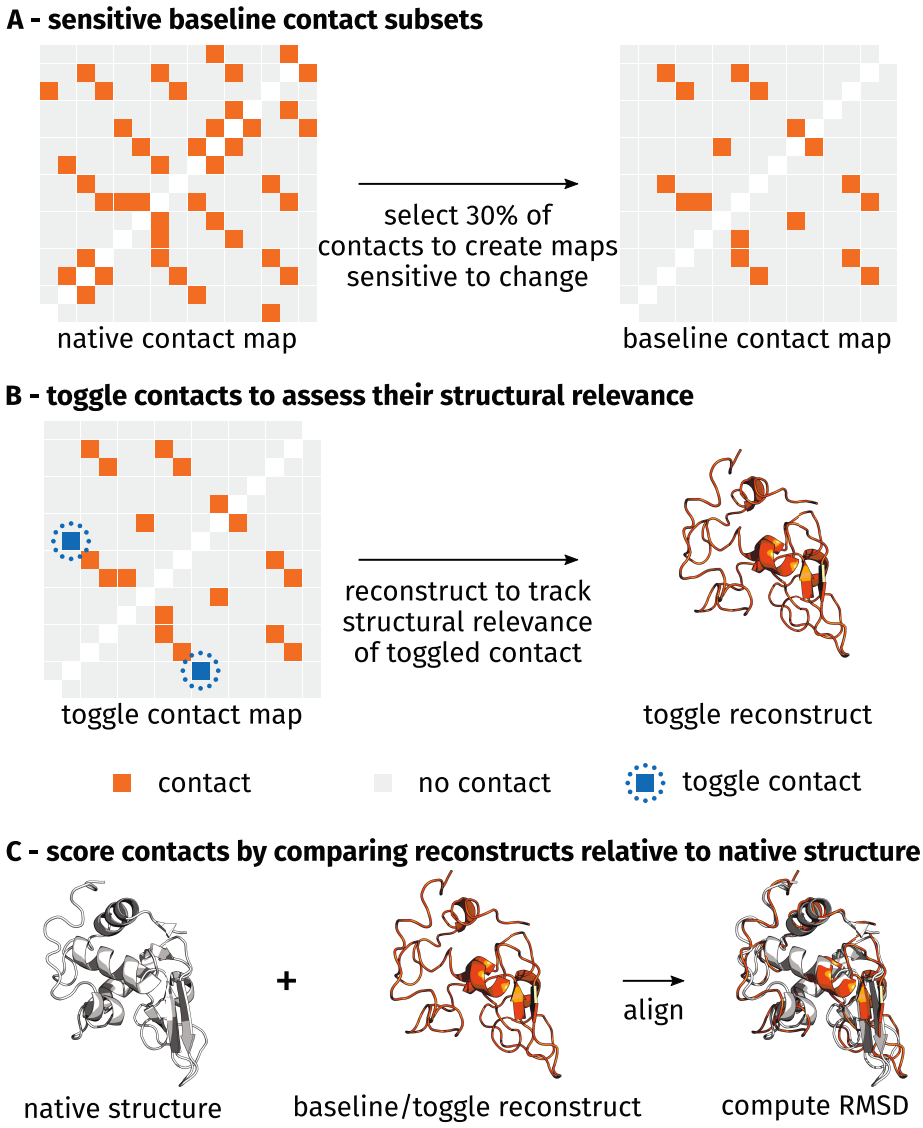

Figure 4.2.: Graphical Depiction of the StructureDistiller Algorithm

In order to assess the structural relevance of individual contacts, the effect of their consideration on the reconstruction performance (ΔRMSD) is computed. This allows a novel, more fine-grained interpretation of CMs. (**A**) By using 30% of all contacts present in the native CM, baseline CMs are created which provide maximum sensitivity to the removal or addition of a single contact. (**B**) Within these baseline CMs, all contacts of the native CM are toggled: contacts already present are removed and those absent are added. Reconstructs are created based on these toggle CMs. (**C**) By superimposing reconstruct and native structure, the structural relevance of all contacts can be quantified as relative change in RMSD.

up to 200 residues requires one day on average.

4.1.5. DEFINITION OF RECONSTRUCTION STRATEGIES

Various reconstruction strategies were used to assess the relevance of contacts in a CM. In all cases, a number equal to 30% of the contact count in the native map was used. For the creation of the random bin, 30% of all native contacts were chosen randomly. Best constitutes the 30% of all contacts sorted for highest structural relevance, worst resembles 30% of all contacts with the lowest structural relevance. All percentage numbers are relative to the number of contacts in the native structure. All operations on all definitions

Algorithm 1 StructureDistiller Pseudocode

procedure SD(native structure S_{native}, redundancy r, coverage v)
    ▷ initialization
    create set of contacts C_{native} using S_{native}
    ▷ create r baseline reconstructions
    for $i = 0 : r$ **do**
        create sampled subset $C_{baseline,i}$ of C_{native} with coverage v
        reconstruct structure $S_{baseline,i}$ from $C_{baseline,i}$
        superimpose S_{native} and $S_{baseline,i}$
        measure performance by $RMSD_{baseline,i}$
    end for
    ▷ toggle all contacts in baseline reconstructions
    for $c \in C_{native}$ **do**
        for $i = 0 : r$ **do**
            if $c \in C_{baseline,i}$ **then**
                create toggle subset $C_{toggle,i}$ by removing c from $C_{baseline,i}$
            else
                create toggle subset $C_{toggle,i}$ by adding c to $C_{baseline,i}$
            end if
            reconstruct structure $S_{toggle,i}$ from $C_{toggle,i}$
            superimpose S_{native} and $S_{toggle,i}$
            measure performance by $RMSD_{toggle,i}$
            ▷ compute structural relevance of contact c
            if $c \in C_{baseline,i}$ **then**
                $\Delta RMSD_c = RMSD_{baseline,i} - RMSD_{toggle,i}$
            else
                $\Delta RMSD_c = RMSD_{toggle,i} - RMSD_{baseline,i}$
            end if
        end for
    end for
    return set of all $\Delta RMSD_c$
end procedure

are performed with ten-fold redundancy. Contact distances were assessed: all short (sequence separation of 6–11) and long (>23) contacts [169] were assessed. The same was done for hydrogen bonds and hydrophobic interactions. Because the number of contacts of a particular distance or type may be small, a dedicated bin (e.g. non-short) was created to match in size.

4.1.6. INTRODUCTION OF FALSE POSITIVE CONTACTS

False positive contacts are contacts not present in the contact map of the native protein structure. CMs were created by the best and random strategy and in 1% bins up to 10% false positive contacts were introduced, replacing the initially selected native contacts. Analogous to the employed contact definition, false positive contacts were required to exhibit a sequence separation greater than five.

4.1.7. STATISTICAL ANALYSIS

Residues without any contacts (i.e. where no structural relevance can be computed) were ignored from statistical analysis. Notched box plots were used for visualization. The notch

corresponds to the 95% confidence interval around the median. When the notches of two distributions do not overlap, they can be assumed to be different. Significance was explicitly tested by the Mann-Whitney U test.

4.2. RESULTS & DISCUSSION

A subset of the Start2Fold dataset [8] encompassing 30 proteins was used for further analysis, thereby only proteins with an annotation of EFR were considered. The folding and stability characteristics of the corresponding proteins have been determined by HDX experiments [105, 8, 11] and these properties may relate to the most relevant contacts of a CM and constitute a direct connection of protein folding *in vivo* and structure prediction *in silico*. Individual contacts cannot be directly assessed regarding their structural relevance (i.e. how much does knowledge of this contact improve reconstruction) because a single contact will never yield a meaningful reconstruct, instead they depend on a set of other contacts [14, 15]. In order to quantify the structural relevance of individual contacts, they have to be disentangled from all other contacts mandatory for a meaningful reconstruction in the first place (see Section 4.1.4). The reconstruction error describes the dissimilarity of each reconstruct with respect to the native structure [14, 15].

For a basic assessment, all proteins were reduced to a CM representation and random subsets with varying coverage were used to reconstruct all proteins (Figure 4.3). With increasing number of considered contacts, the reconstruction error decreases. The reconstruction process using more contacts becomes more robust as the distributions decrease in variance. At 30% coverage of all native contacts the yielded reconstructs resemble the fold of the native structure and are also sensitive to the removal or addition of individual contacts. The reconstruction error approaches 2 Å when all contacts are used as described in literature [15].

4.2.1. THE STRUCTURAL RELEVANCE OF INDIVIDUAL CONTACTS AND RESIDUES

The structural relevance of 5,173 contacts was quantified by the StructureDistiller algorithm (see Section 4.1.4). The outputted score captures the average performance increase in Å (called ΔRMSD), when a particular contact is considered for the reconstruction process compared to a reconstruction without knowledge of this particular contact. Positive structural relevance scores indicate contacts which favorably contribute to reconstruction fidelity, whereas negative scores indicate native contacts which hinder or at least not substantially improve the process. The removal of contacts results in a structural relevance of 0.012 ± 0.253 Å. In contrast, the addition of a contact leads to 0.022 ± 0.253 Å. Most contacts contribute positively to reconstruction performance. An increased change in reconstruction quality can be observed for the removal of contacts because the relative change in number of considered contacts is greater compared to the addition of contacts. Only a small number of contacts is of high structural relevance with similar tendencies shown by studies on CMs [14, 15] as well as protein folding in general [170], where good reconstructions as well as correctly folded protein structures depend on a small number of key contacts. The high variance of the structural relevance scores is the result of both the CM sampling as well as the reconstruction routine [114] being stochastic processes. Both operations are performed redundantly to address this issue.

Several features (Table 4.1) are used to describe contacts in more detail. Their relation to the structural relevance score was assessed. Therefore, residue contacts are distinguished according to their sequence separation [169]. Short contacts (6–11) exhibit a significant decrease in structural relevance. In contrast, long contacts (>23) of sequentially highly separated residues are more common and feature increased structural relevance scores. The

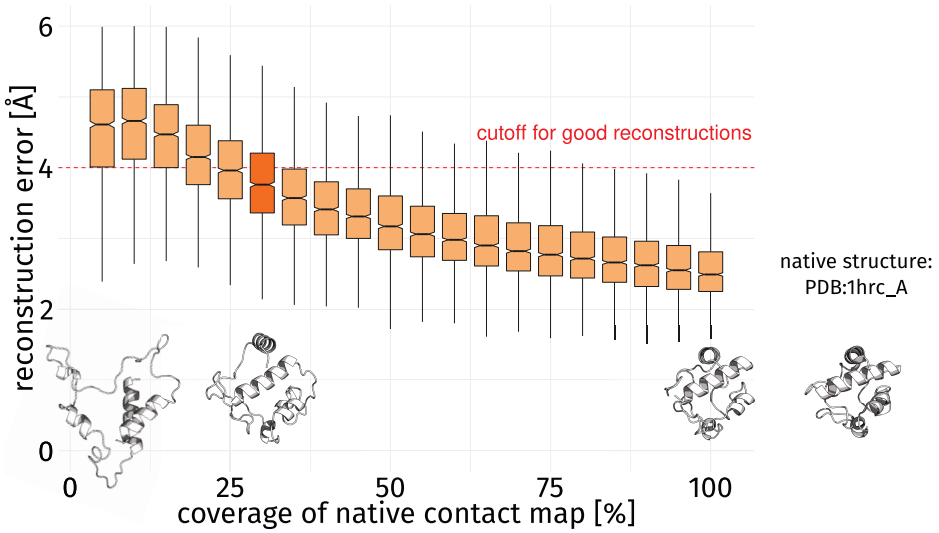

Figure 4.3.: Reconstruction Error by Percentage of Contacts

When more contacts are considered, the average reconstruction error decreases [14] and the same is true for the variance of each bin. For the assessment of the structural relevance of contacts, 30% of all native contacts (box plot filled dark orange) were chosen as compromise because it ensures reconstructs of average quality while the corresponding CMs are still sensitive to the removal or addition of individual contacts (as indicated by a big shift in reconstruction error with respect to the neighboring bins). Thus, 4 Å was used as cutoff for good reconstructions for further analyses. Renderings of four structures are provided to make the influence of the coverage of the native CM more tangible. They resemble knowledge of 5%, 30%, and 100% of all native contacts as well as the native structure (PDB:1hrc_A).

change is insignificant for contacts of medium range. Non-covalent interactions such as hydrogen bonds and hydrophobic interactions are annotated between residues [37] and contacts which are backed by either contact type are considered. Both are associated with a significant change whereby the presence of hydrogen bonds decreases the structural relevance and hydrophobic interactions increase the structural relevance of a contact.

Previously it has been shown that contacts within as well as between secondary structure elements are required for optimal reconstruction performance [14, 15]. Commonly, reconstructions only consider residue pairs at least five positions apart on sequence level [169], though there are cases where the usually ignored contacts may contribute valuable information on the structure of loops [171]. Non-covalent contacts have a significant effect on the structural relevance of a contact. Hydrogen bonds occur between backbone atoms of amino acids where they define and stabilize secondary structure elements. Some amino acids such as serine or threonine feature polar side chains which allow them to engage more flexibly in this type of non-covalent contact. The importance of hydrogen bonds furnished by side chains for protein folding and stability has been shown [41, 27]. Hydrogen bonds may feature lower structural relevance scores because of their propensity to occur between polar amino acids at positions exposed to the solvent. In contrast, hydrophobic interactions primarily occur in the buried hydrophobic core of a protein where they are surrounded by many other residues which reduces the degree of freedom. Especially, the importance of tertiary contacts furnished by hydrophobic interactions has been shown [87]. Such contacts provide information on the correct assembly of distant parts of the protein and, thus, are relevant for structural integrity both during protein folding and in structure prediction.

Regarding contact prediction methods, evolutionary couplings [140, 141] show no significant association to the structural relevance score. However, a slight increase in structural relevance can be observed, when two positions are evolutionarily coupled. A selection of

Table 4.1.: Contact-Level Features Influencing the Structural Relevance Score

feature	category	n	μ	σ	trend	p-value
short contact (6–11)	no	4,053	0.022	0.086	▽	0.025
	yes	1,120	0.014	0.081		
medium contact (12–23)	no	3,902	0.021	0.085	–	0.161
	yes	1,271	0.018	0.086		
long contact (>23)	no	2,391	0.016	0.084	△	0.002
	yes	2,782	0.024	0.086		
hydrogen bond	no	4,610	0.021	0.085	▽	0.018
	yes	563	0.011	0.083		
hydrophobic interaction	no	4,632	0.019	0.084	△	<0.001
	yes	541	0.029	0.095		
evolutionarily coupling	no	3,246	0.018	0.087	–	0.203
	yes	1,461	0.022	0.084		
top-scoring coupling	no	3,687	0.018	0.087	–	0.059
	yes	1,020	0.024	0.083		

Contact length refers to the sequence separation of the contact [169]. Hydrogen bond and hydrophobic interaction refers to contacts for which the respective interaction type was observed [37]. Evolutionary couplings by DCA [140, 141], for some proteins no data could be computed. Top-scoring couplings are the first $0.4L$ contacts sorted by their coupling rank. n describes the number of observations, μ the corresponding average, and σ the respective standard deviation. The trend is given, i.e. does presence of this feature decrease (▽) or increase (△) the structural relevance scores. Insignificant change is represented by a dash (–).

the $0.4L$ top-scoring contacts (L refers to the sequence length) results in a more substantial, though still insignificant, change in structural relevance. Many predicted couplings are not actually present in the native CM despite being meaningful for structure reconstruction. Also, potential false positive predictions by the sequence co-variation techniques such as the DCA are not evaluated, which can be expected to have a negative effect on reconstruction quality [17].

At residue level, a set of features was evaluated with the same reasoning (Table 4.2). Residues in unordered secondary structure elements have significantly lower structural relevance than those in α-helices and β-strands. For ordered secondary structure elements, backbone angles and hydrogen bonding patterns are used as additional constraints during reconstruction [114] which may explain an overall performance increase. The previous association of hydrophobic interactions and structural relevance may be explained by a bias for buried residues; however, no significant association is observed at residue level. The annotation of EFR does not influence structural relevance significantly, while the opposite is true for HSR (see below). Functional residues may not be of structural relevance, because binding sites tend to be exposed to the solvent and commonly have unfavorable conformations [172]. Again evolutionary couplings [140, 141] do not lead to increased structural relevance, probably because most residues feature at least one predicted coupling. Filtering for the $0.4L$ top-scoring positions (i.e. regarding their cumulative coupling strength) does not lead to a significant change either.

4.2.2. MOST RELEVANT CONTACTS INCREASE RECONSTRUCTION PERFORMANCE AND RESILIENCE TO FALSE POSITIVE PREDICTIONS

The subset of contacts with high structural relevance should lead to good reconstructs when combined. Therefore, proteins were reconstructed using various strategies to select sub-

Table 4.2.: Residue-Level Features Influencing the Average Structural Relevance Score

feature	category	n	μ	σ	trend	p-value
coil	no	1,533	0.029	0.060	▽	<0.001
	yes	996	0.019	0.064		
buried	no	1,424	0.024	0.067	–	0.075
	yes	1,105	0.027	0.055		
early folding	no	2,115	0.025	0.062	–	0.543
	yes	414	0.026	0.061		
highly stable	no	1,731	0.021	0.061	△	<0.001
	yes	688	0.030	0.062		
functional	no	2,078	0.026	0.059	–	0.919
	yes	119	0.028	0.062		
evolutionarily coupled	no	503	0.026	0.064	–	0.754
	yes	1,975	0.026	0.062		
top-scoring coupled	no	1,361	0.025	0.066	–	0.492
	yes	1,117	0.026	0.057		

Residues in coil secondary structure elements and residues buried according to their RASA were evaluated. Residues were assessed regarding their folding characteristics [8]. Annotation of functional residues from UniProt [173]. Considers evolutionary couplings and the $0.4L$ top-scoring positions according to the cumulative coupling strength [140, 141]. n describes the number of observations, μ the corresponding average, and σ the respective standard deviation. The trend is given, i.e. does presence of this feature decrease (▽) or increase (△) the structural relevance scores. Insignificant change is represented by a dash (–).

sets equal to 30% of all native contacts (Figure 4.4). To provide a baseline, 30% of randomly selected contacts were considered (gray). Both strategies are based on the computed structural relevance scores of the corresponding proteins. Therefore, the contact list was sorted by descending structural relevance scores. The 30% most relevant contacts were selected (green). In contrast, the 30% least relevant contacts constitute another reconstruction strategy (red). Other interesting aspects are contact distance and type: therefore short (6–11), long (>23) contacts, hydrogen bonds, and hydrophobic interactions were assessed.

The RMSD is used to quantify the fidelity of a reconstruct by aligning it to the native structure – high reconstruction errors occur for bad reconstructs. A random selection of 30% of contacts achieves 3.839 ± 0.599 Å. A combination of contacts by the most relevant strategy significantly outperforms the random strategy with an average reconstruction error of 3.479 ± 0.625 Å. Consideration of the least relevant scoring contacts results in an increase in reconstruction error to 4.311 ± 0.687 Å.

Chen et al. assumed that no rational selection of contacts can surpass a random selection in terms of reconstruction fidelity [14]. Later, Sathyapriya and coworkers [15] provided an algorithm capable of doing just that. It is especially remarkable that their approach merely evaluates which neighborhood is shared by a pair of residues. The main aspect of their algorithm is the selection of non-redundant contacts which can provide the maximum amount of information for a reconstruction when combined. The selection of the most relevant contacts as determined by StructureDistiller constitutes a different approach to compose a set of contacts which allow for better reconstructs than a random selection. Of all native contacts two selections can be readily made. One is significantly better suited for reconstruction purposes than a random selection and whereas the other one performs significantly worse. It is also remarkable that a combination of long contacts performs significantly worse than the negated selection, despite individual long contacts exhibiting high structural relevance (Table 4.1). This emphasizes the context-specificity of individual con-

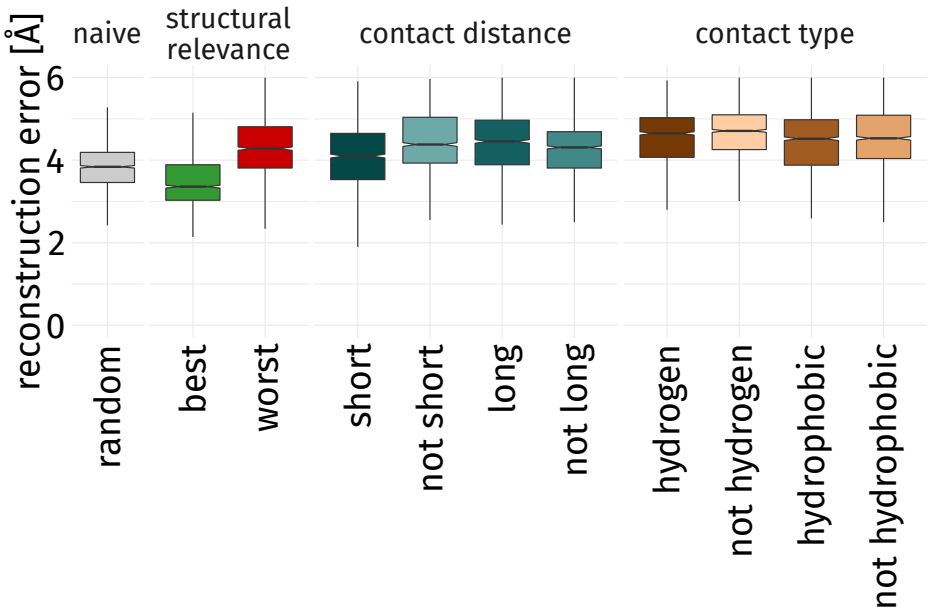

Figure 4.4.: Impact of Reconstruction Strategy on Performance

Various strategies were used to reconstruct structures of the dataset using a number of constraints equal to 30% of contacts in the native map. A random selection of contacts (gray), the most relevant ones by structural relevance (green), and the least relevant ones (red). The most relevant contacts yield the lowest reconstruction error when combined. This configuration outperforms a random selection of contacts significantly. Previous studies [14, 15] have shown the difficulties in finding combinations of contacts yielding better reconstructs than a random selection. Using the least relevant contacts results in an increased error compared to the random selection. When only a subset of all entries of a CM can be considered (as it is commonly the case [117] and reasonable for efficiency [15]), it is crucial for reconstruction performance which subset of contacts is chosen. Contact distance and type bins are only comparable to the explicitly negated bins because the available number of contacts differs (i.e. there may not be enough hydrogen bonds to match the number of contacts in the random bin). Not selections based on contact distance or type perform worse than or comparable to their counterpart which implies the necessity to consider a complex collection of contacts for a successful reconstruction [14]. When combined short contacts yield relatively good reconstructs even though they structural relevance scores are low.

tacts and substantiates the previous findings [14], wherein both short and long contacts are needed for good reconstructions.

The sensitivity of a CM to false positive contacts has been discussed before – even a small number of such contacts not present in the native structure is detrimental to reconstruction performance [17]. As shown previously, contacts with high structural relevance allow for better reconstructs. Interestingly, the selection of the most relevant contacts also can compensate the moderate introduction of false positive contacts (Figure 4.5). The selection of the most relevant contacts performs significantly better than a random selection in all considered cases. The introduction of false positive contacts quickly leads to reconstructions with errors above 4 Å as larger fractions of false positive contacts dilute the correct information captured by native contacts. When more than 7% false positive contacts are introduced to the best bin, the majority of reconstructions is of bad quality. When 30% of all contacts are selected randomly, only 3% false positive contacts can be introduced before the error exceeds the threshold of 4 Å. The consideration of the most relevant contacts buffers the negative influence of false positive predictions (Table 4.3): median performance

is comparable between reconstructions based on a random selection without false positive contacts and the selection of the best contacts diluted by 6% of false positive contacts.

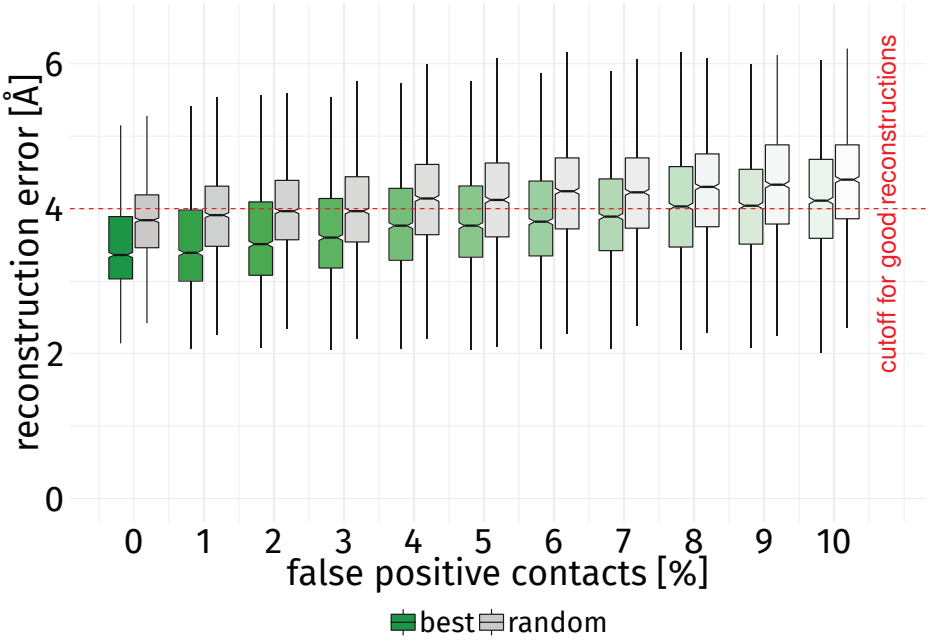

Figure 4.5.: Influence of False Positive Contacts

The reconstruction error is given of for 30% of all contacts in the best (green) and random (gray) bins with an increasing fraction of false positive contacts. In all cases, the most relevant contacts perform significantly better than a random selection when it comes to compensating false positive contacts (p-value <0.001). E.g., the median performance of a random selection without false positive contacts is comparable to that of the best selection with 6% false positive contacts. When more than 3% false positive contacts are introduced into the random selection, the error of the majority of reconstructions lies above 4 Å, whereas the best selection can compensate more than double the number of false positive contacts before surpassing this threshold. Knowledge of the most relevant contacts in a protein structure thus increases the resilience to false positive contacts as well as the overall reconstruction performance.

This implies that contact prediction methods should not only minimize the number of false positive predictions but may benefit from ignoring particular native contacts when they exhibit low structural relevance scores. Since even those native contacts can hinder reconstruction (as indicated by negative structural relevance scores), it becomes evident that the correct ranking of contacts [114, 148] has a serious influence on reconstruction quality and should be considered for the design and training of contact prediction techniques. Potentially, increased structural relevance and capabilities to compensate false positive predictions (or mutations) cannot only be observed for *in silico* reconstruction, but may be also influence the *in vivo* folding process. The insignificant association of evolutionary couplings and structural relevance scores suggests that the most relevant contacts may not be easy to predict but can contribute significantly more information needed for the successful reconstruction of a protein.

4.2.3. ANALYSIS OF EARLY FOLDING AND HIGHLY STABLE RESIDUES

A direct connection to particular folding and stability characteristics is provided by the annotation of EFR which initiate and guide the folding process. However, according to the structural relevance score no change can be observed for EFR (Table 4.2). Contacts of HSR exhibit a significant increase in structural relevance compared to unstable contacts [8]. It

Table 4.3.: Reconstruction Error Introduced by False Positive Contacts

false positive contacts [%]	μ_{best}	$\widetilde{x}_{\text{best}}$	μ_{random}	$\widetilde{x}_{\text{random}}$
0	3.479	3.360	3.839	3.840
1	3.498	3.390	3.903	3.910
2	3.598	3.510	3.971	3.965
3	3.665	3.600	3.996	3.965
4	3.808	3.765	4.135	4.140
5	3.840	3.765	4.117	4.120
6	3.882	3.820	4.211	4.240
7	3.931	3.890	4.229	4.225
8	4.028	4.030	4.256	4.300
9	4.038	4.040	4.292	4.330
10	4.140	4.110	4.354	4.400

For increasing rates of false positive contacts the reconstruction performance using 30% of the native contacts are given. μ_{best} refers to the average performance using the most relevant contacts, μ_{random} to that using a random selection of contacts. $\widetilde{x}$ describes the median of the corresponding population. In all cases, the performance of the best bin is significantly better than that of a random selection.

is remarkable that contacts of EFR show no increase in structural relevance despite their presumed role for the protein folding process [11, 174]. A possible interpretation is that EFR primarily define stable, local structures [11, 174] due to their occurrence in sequence regions associated to high backbone rigidity [20]. They form defined sequence regions with fewer possible backbone conformations and produce pivotal secondary structure elements [20]. Therefore, EFR define the folding nucleus of a protein and sequentially encode the ordered secondary structure elements formed first. However, crucial contacts between these secondary structure elements may be mediated by other residues which are not necessarily EFR themselves but may occur in secondary structure elements containing EFR [20].

Another aspect of this data is the annotation of residues which are strongly protected in stability measurements [8]. Such residues occur in stable, local structures and their contacts are beneficial to reconstruction performance. EFR constitute more rigid backbone fragments; thus, these residues may occur in regions which lowered conformational entropy. The decreased flexibility may enforce specific contacts between distant protein regions which need to form contacts crucial for the formation and stabilization of the native structure. The importance of ordered secondary structure elements with above average stability has been described e.g. for the assembly of helices [175].

The defined-pathway model [176, 18, 19] describes protein folding as a deterministic, hierarchic process. EFR occur in regions which autonomously fold first relative to the rest of a protein. Furthermore, this tendency does not depend on tertiary contacts in a protein structure, but is rather the direct consequence of the local sequence composition [20, 13, 11]. These stable, local structures may be secondary structure elements [13] or larger autonomously folding units also referred to as foldons [18]. In a stepwise process, such local structures will subsequently establish tertiary contacts and assemble the native conformation of a protein [82, 177, 18]. The employed reconstruction method directly considers secondary structure elements, which are used as additional constraints. Therefore, most secondary structure elements should be represented successfully which may explain why long contacts are particularly important for structural integrity. It is also reasonable that the structural relevance of a contact increases with the distance at the sequence level, because such constraints do not only enforce the correct placement of both residues but also have an indirect positive impact on the correct conformation of residues in between.

4.2.4. DISRUPTION TO CYTOCHROME C INDUCES MOLTEN GLOBULE STATE

Ground truth on the structural importance of individual contacts is difficult to find – as a case study cytochrome c is used which is also element of the dataset. Cytochrome c (Figure 4.6) contains two Ω-loops which are stabilized by a hydrogen bond between HIS-26 and PRO-44. This contact is considered structurally relevant, because disruptions to it have been shown to induce a molten globule state [178, 179]. Particularized folding studies [105] have also identified the N- and C-terminal helices as foldons, i.e. autonomously folding units which initiate and guide the folding process. Besides that, wide parts of the structure are constituted of coil regions and fixate a heme ligand, thus potentially exhibiting increased structural flexibility.

The computed structural relevance of many residues of cytochrome c is neutral or even negative. Especially coil regions feature contacts which tend to decrease reconstruction fidelity. Remarkable are the high structural relevance scores of HIS-26 and GLY-45 as well as their direct contact for which the score amounts to 0.172 Å (rendering it the fifth most relevant contact). No structural relevance is reported for PRO-44 as it does not participate in any contacts according to the employed contact definition, though both groups are positioned in a way which would allow them to form a hydrogen bond. In literature [178], the contact between HIS-26 and PRO-44 is reported as crucial for the correct conformation of cytochrome c. Disruptions will result in a loss of structure [178], though the relevance of PRO-44 may also be attributed to the backbone rigidity introduced by the proline residue. The detection of relevant contacts and positions is fuzzy [29], but the high scoring contact between HIS-26 and GLY-45 implies the importance of a contact between both Ω-loops for successful protein folding as well as structure reconstruction. Between GLY-29 and MET-80 the most relevant contact is located which increases reconstruction fidelity by 0.563 Å on average. This contact occurs between two unordered coil regions as well and implying that some structural information on the correct arrangement of these unordered protein parts is crucial for a successful reconstruction. Mutations to HIS-33 have been demonstrated to show no effect [178] which is also captured by slightly negative structural relevance score of -0.015 Å. Both N- and C-terminal helix contain residues with high relevance, especially in regions where both helices interact. The importance of these helix contacts has been shown previously [180]. The role of both helices as foldons [105] points to a high intrinsic stability.

The structural relevance score successfully spots contacts and residues crucial for structure integrity as shown in experiments [178, 105, 180]. The previously described contact between HIS-26 and PRO-44 [178] is absent as the result of a too strict contact definition, yet the necessity of structural information in this region is captured nevertheless.

4.3. CONCLUSION

CMs are one of the most prominent tools in today's structural bioinformatics [124, 139], though mere knowledge of residue contacts can neither describe all events of the protein folding process [175] nor is it the optimal basis of structure prediction techniques [123]. It is demonstrated that native contacts in a protein structure are not of equal importance for the reconstruction of the tertiary structure from this reduced representation. A more fine-grained interpretation of CMs can be achieved by the consideration of structural relevance scores. Contacts of high structural relevance tend to be unique contacts for which no backup exists as it is the case for the contact between two Ω-loops in cytochrome c [178]. The importance of this contact for the structural integrity also implies that high structural relevance scores may capture crucial positions for structure stability. Another implication of high scoring contacts lies in fold recognition as they may be the footprint of particular pro-

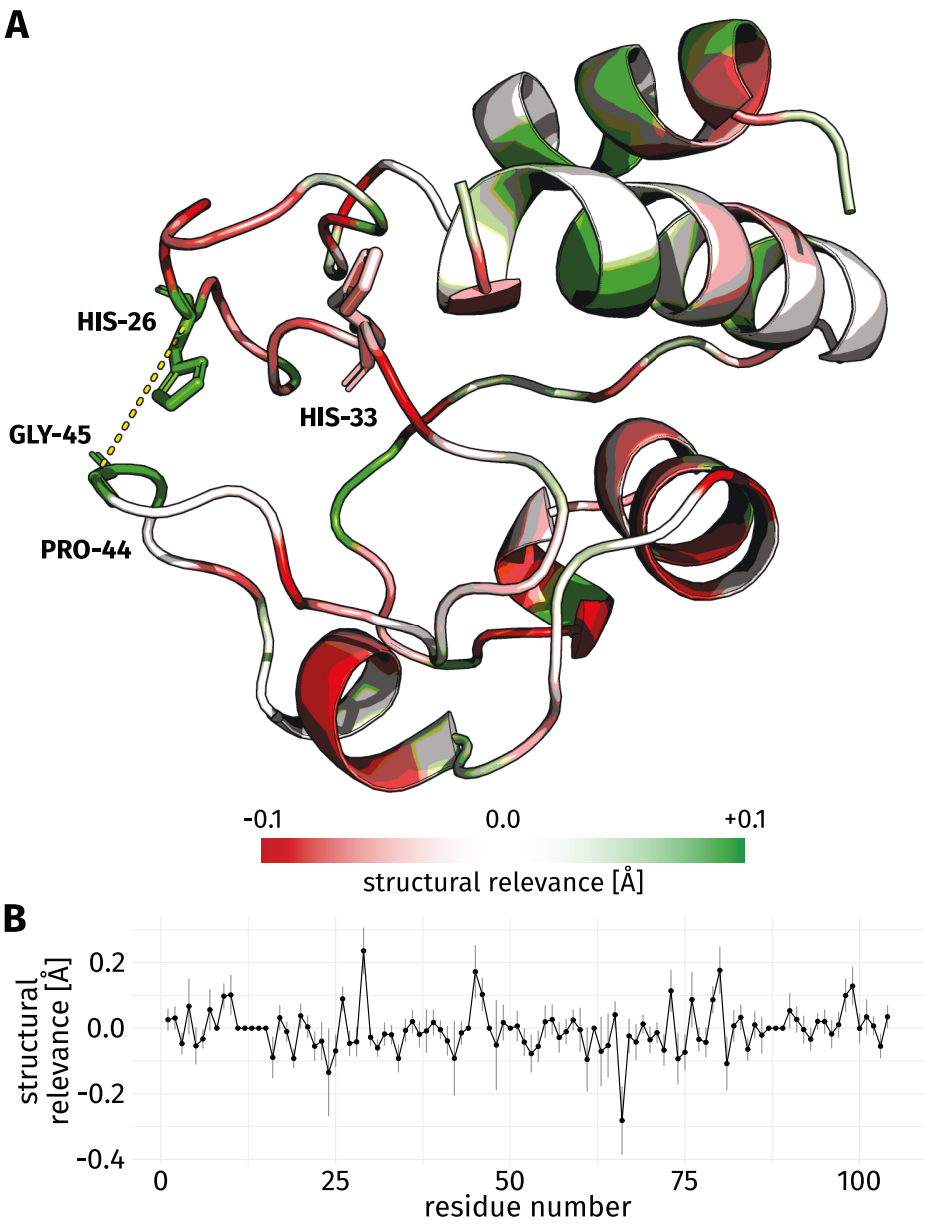

Figure 4.6.: Cytochrome C Colored by Structural Relevance

(**A**) Depicted is PDB:1hrc_A. Residues with high structural relevance are depicted in green, those with negative structural relevance are rendered in red. For gray residues no contacts were observed, and no structural relevance scores are reported. It has been shown in experiment that disruptions to the hydrogen bond between HIS-26 and PRO-44 will induce a molten globule state when the association between both Ω-loops is disrupted [178, 179]. StructureDistiller reports high relevance for HIS-26, GLY-45, and the contact both share (yellow dashed line), though no direct contact is detected between HIS-26 and PRO-44 due to the employed contact definition. HIS-33 has been described as variable position lacking any structurally relevant contacts [178] and this observation is manifested in the low structural relevance score of this residue. The N- and C-terminal helices exhibit high structural relevance, especially for residues which constitute their interface. Both helices have been shown to be foldons which initiate the folding process of cytochrome c [105]. Other parts of the structure are primarily composed by coil regions, fixate a heme ligand, and show low structural relevance. (**B**) Per residue structural relevance as line chart. The standard deviation is given for each point. Residues without contacts exhibit a relevance of 0 Å.

tein folds.

Coevolution or SML techniques are the basis for the prediction of CMs [141, 124, 135]. Conventionally, contact predictors are designed and trained on collections of all native contacts in a dataset. Subsequently, the most reliable contacts are selected from all predictions; the size of this subset depends on sequence length [117]. This study shows that these subsets drastically change in meaningfulness as indicated by reconstruction fidelity. An implication is that it is not the optimal strategy to consider a random subset of contacts, but that prediction accuracy and information per contact may be increased when explicitly the contacts with the highest structural relevance scores are considered. Especially, coevolution techniques struggle with transitive predictions: if there is a signal between residue A and B as well as B and C, then it is likely that for the pair A and C a signal is reported too [140, 141]. Machine learning also comes at a price: an increase of true positive predictions involves more false positives and the consideration of all contacts may lead to a decrease in generalizability. Contact maps and reconstruction algorithms can compensate some false positive predictions [17], but errors in prediction methods tend to be correlated and hinder reconstruction more than random noise [112]. Regardless of the particular approach, it may be beneficial when predictors do not try to cover all native contacts but focus only on the most relevant ones. This would decrease the number of predicted contacts but may increase the reliability of their prediction by avoiding both false positive predictions and emphasizing contacts which promise to improve reconstruction fidelity the most while ignoring those which contribute only marginally.

Residues in a protein are covalently bound and constraints on a residue will also affect neighboring residues. Thus, the mapping of complex information as the structural relevance to individual contacts should be considered with a grain of salt [29]. One of the most delicate aspects when handling CMs is the used contact definition [16, 123]. Particularly, the distance-based contact definition employed in this study does not imply chemically relevant contacts between atoms (such as hydrogen bonds or hydrophobic interactions). The chosen cutoff is rather rigorous and will ignore some meaningful contacts; a relaxation of this cutoff will encompass more contacts but also increases computation time. In some cases such as the assembly of helices [181], information of the complex hydrogen bond patterns has to be considered, which may benefit from a more fine-grained contact definition. Another refinement of the proposed protocol is the explicit consideration of residues which are not in contact. This information has been demonstrated to increase reconstruction fidelity [123]. Also, it is natural that the employed reconstruction pipeline [114] as well as scoring scheme [136] have an effect on the computed scores and may introduce some form of bias to the proposed scores. The TM-score may be more suited to score reconstructs because it is independent of protein length and can more intuitively state whether the correct protein fold was reconstructed [137]. TM-score values are provided as output, but presented results use the RMSD value because of its widespread use and comparable results in relation to previous studies [14, 15, 16]. Furthermore, the decision to use 30% of all native contacts to assess the structural relevance may be not generally applicable. The StructureDistiller algorithm may be improved by determining for each protein structure individually where the sweet spot lies between meaningful reconstructs and maximized sensitivity.

In summary, a way to assess the structural relevance of individual contacts and residues is presented. This constitutes a substantial broadening of the toolkit available for the interpretation of CMs and protein structures in general, while making the connection of CMs and tertiary structure more concrete. Results of the StructureDistiller algorithm may enhance contact prediction techniques [140, 139, 121, 135], CM evaluation [117, 148], reconstruction algorithms [114], quality assessment programs for the yielded models [182], statistical potentials, and may even give particularized insights into the protein folding problem itself [15].

On a general level, the dataset of EFR and HSR [8] provides valuable information to converge on the protein folding problem [163, 13]. The Start2Fold dataset [8] enables the direct connection of protein folding and structure prediction which is furnished by CM representations. It is implied that EFR may initiate protein folding and determine the order in which local structures are assembled [18, 19] but they are of average relevance in terms of structural relevance. HSR may not fold early but constitute regions of a protein which prevent spontaneous unfolding. Interestingly, regions of HSR are of high relevance for the formation and stabilization of the correct protein fold. Furthermore, hydrophobic interactions, contacts of ordered secondary structure elements, as well as long-range contacts promote structural integrity.

Maybe the protein folding problem is not solvable without understanding how protein structures can be predicted reliably. Indeed, both problems are often described to be two sides of the same coin [27] and structure prediction did provide new insights into the folding process before [21, 57]. Additional tools are needed to make the connection of protein sequence and structure more tangible and StructureDistiller provides just that. The algorithm allows a novel fine-grained interpretation of contact maps and may improve their interpretability. Applications of the proposed algorithm are not limited to the Start2Fold database [8], it can be used for the analysis of arbitrary protein structures, e.g. to assess structural effects of mutations at certain residue positions. Following a new paradigm, the interface between protein folding and structure prediction [27] may be explored in detail.

5. CHARACTERIZING THE RELATION OF FUNCTIONAL AND EARLY FOLDING RESIDUES IN PROTEIN STRUCTURES

This Chapter is Based on the Publication:

Bittrich, S., Schroeder, M., & Labudde, D. (2018). Characterizing the relation of functional and Early Folding Residues in protein structures using the example of aminoacyl-tRNA synthetases. *PloS one*. 13(10), e206369.

This Chapter Employs Methods from:

Bittrich, S., Heinke, F., & Labudde, D. (2016). eQuant – A Server for Fast Protein Model Quality Assessment by Integrating High-Dimensional Data and Machine Learning. In *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery* (pp. 419-433). Springer, Cham.

The events of the earliest stages of folding are enigmatic [102, 103, 72]. EFR are the residues which fold first during the folding process and are a resource to address an integral question: are the experimental signals of EFR transient implying that EFR are only relevant in the early stages of the folding process or will EFR also exhibit distinct characteristics in the successfully folded, native conformation?

It is unknown what sequence features causes particular residues to fold early and how these residues contribute to the formation of the native structure (Figure 5.1A). EFR are connected to the defined-pathway model and provide an opportunity to understand the driving forces behind the formation of stabilizing local structures as well as the formation of tertiary contacts [18, 13].

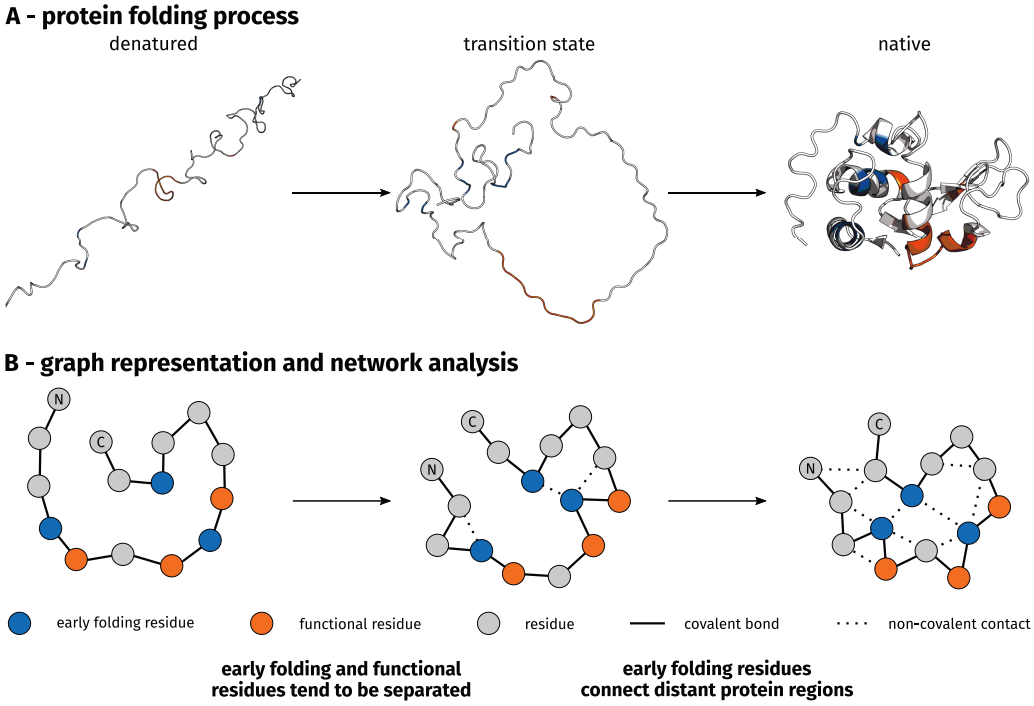

Figure 5.1.: The Relation of Early Folding and Functional Residues

(A) During the folding process, an extended protein chain passes the transition state and forms a native structure [28]. (B) Protein structures are represented as graphs to derive topological descriptors of residues. Amino acids constitute nodes, whereas residue contacts are represented as edges. EFR are structurally relevant residues which participate early in the folding process by forming local contacts to other residues. They are separated from functional residues which are primarily ligand binding sites and active sites as derived from UniProt [173]. EFR show a great number of tertiary contacts which furnish the spatial arrangement of protein parts despite being far apart at sequence level.

Several novel structural features are employed for the characterization of EFR. Especially, the Energy Profiling approach, topological descriptors based on residue graphs, and the explicit consideration of non-covalent contact types provide a new level of information in order to describe the folding process. EFR exhibit lower, more stable computed energies in their Energy Profile [183, 182]. Network analysis reveals that EFR are more connected to other residues and that they are located at crucial positions in the residue graph (Figure 5.1B). This distinct wiring to the rest of the protein is especially furnished by hydrophobic interactions. EFR are likely structurally relevant for the correct protein fold [11]. This information is used to demonstrate that many proteins separate structurally relevant residues from functional residues (Figure 5.1B).

5.1. MATERIALS & METHODS

5.1.1. DATASET CREATION

Folding characteristics of residues were obtained from the Start2Fold database [8]. Therein, the authors adopted the definition of EFR from Li et al. [87] and presented a refined dataset which ignores possible back-unfolding and aggregation events [184]. The database covers all structural protein families present in the CATH and SCOP databases [11]. However, the size of the deposited proteins [8, 11] varies from 56 to 164 residues (Table 5.1) which makes this resource primarily relevant for the folding of similarly small proteins. Because local sequence features determine where EFR are located, this characteristic may be independent of sequence length and applicable for a wider range of proteins. The original dataset contains two groups of similar sequences of lysozymes (PDB:1hel_A, PDB:1lz1_A, PDB:2eql_A) and apo-myoglobins (PDB:1mbc_A, PDB:1ymb_A). In these cases, we only considered the structure with the highest resolution.

This procedure resulted in a dataset for EFR characteristics encompassing 27 proteins and 2,966 residues – 450 (15.2%) of the EFR class and 2,516 (84.8%) of the LFR class. Due to the nature of the HDX experiments no data can be obtained for proline residues which feature no amide group susceptible to HDX [96], rendering them LFR in any case. Annotation of functional residues was performed using the structure integration with function, taxonomy and sequences resource (SIFTS) [185] and UniProt [173] resource. We collected entries in the "Function" and "Family & Domains" section when they were associated to protein function. For 22 proteins an annotation of functional residues existed, totaling in 2,490 residues – 152 (6.1%) classified as functional and 2,338 (93.9%) as non-functional. Residues annotated as functional are summarized in S4 File of the corresponding publication [174] which contains a description of the matched XML tags of functional residues.

5.1.2. GRAPH REPRESENTATION AND ANALYSIS

Protein structures are commonly represented as residue graphs: amino acids constitute the nodes and contacts between residues are represented as edges [186, 187, 188]. This allows a scale-invariant characterization of the neighborhood relation of individual amino acids in the context of the whole protein [187].

In this study, amino acids constitute the nodes of a graph, whereas covalent bonds and residue contacts are represented as edges. Residues were considered in contact when their C_α atoms were less than 8 Å apart. Furthermore, contacts were labeled as either local (i.e. the separation in sequence is less than six) or tertiary (i.e. sequence separation greater than five) [123]. This distinguishes contacts stabilizing secondary structure elements and those which represent contacts between secondary structure elements. The set of distinct neighborhoods of a node is defined as all adjacent nodes which do not share any local edge to any element of the set. Betweenness is defined the number of shortest paths on the graph passing through a specific node, normalized by the number of node pairs [189, 186]. Closeness of a node is defined as the inverse of the average path length to any other node [190]. The clustering coefficient of a node is the number of edges between its n_k adjacent nodes divided by the maximal number of edges between n_k nodes: $0.5 \cdot n_k \cdot (n_k - 1)$ [186]. All topological properties are represented graphically in Figure 5.2.

5.1.3. ENERGY PROFILING

In contrast to the bottom-up approach provided e.g. by quantum chemistry, coarse-grained energy models evaluate datasets of protein structures and apply this knowledge to approximate the energy of other protein structures [191].

Table 5.1.: Early Folding Residue Dataset Summary

Start2Fold	PDB	UniProt	residues	early	functional	int.	shift
STF0001	2abd	P07107	86	39	9	5	1.07%
STF0004	1mbc	P02185	153	12	2	0	-0.10%
STF0006	1am7	P03706	154	29	1	0	-0.12%
STF0008	1coe	P60770	62	6	2	0	-0.31%
STF0009	1a64	P08921	94	19	27	8	2.70%
STF0010	1hce	P13231	118	10	-	-	-
STF0011	5dfr	P0ABQ4	154	5	22	0	-0.46%
STF0012	9pcy	P00287	99	6	4	0	-0.24%
STF0014	1lz1	P61626	130	13	2	0	-0.15%
STF0015	1onc	P22069	104	31	7	2	-0.08%
STF0016	1omu	P68390	56	4	2	0	-0.26%
STF0018	1f21	P0A7Y4	152	16	4	1	0.38%
STF0019	1ygw	P00651	104	13	3	1	0.60%
STF0020	1joo	P00644	149	9	6	1	0.43%
STF0021	2lzm	D9IEF7	164	7	8	1	0.40%
STF0023	1hrc	P00004	104	13	4	0	-0.48%
STF0024	1rg8	P05230	141	38	18	2	-2.02%
STF0025	1e3y	Q13158	104	24	-	-	-
STF0026	1hrh	P03366	125	13	4	0	-0.33%
STF0028	2vil	P02640	126	31	8	0	-1.56%
STF0037	2crt	P60301	60	12	-	-	-
STF0038	5pti	P00974	58	7	2	0	-0.42%
STF0040	1pga	P06654	56	26	-	-	-
STF0043	1i1b	P01584	151	21	6	0	-0.55%
STF0044	2ptl	Q53291	78	12	-	-	-
STF0045	1bdd	P38507	60	20	-	-	-
STF0046	1rbx	P61823	124	14	11	1	-0.20%
$\sum$			2,490	346	152	2	0.04%

Summarizes identifiers [8] of each entry as well as the number of residues in the corresponding protein chain, the number of EFR and functional residues as well as the cardinality of the intersection of both sets.To assess the relevance of the observed intersection it was compared to the expected intersection. Negative shift values occur when the observed intersection is smaller than that expected by the individual frequencies of EFR and functional residues. Positive values are observed when the overlap is more pronounced than to be expected. Proteins not containing any functional residues according to UniProt [173] are marked with dashes. In all cases, chain A of the protein was considered.

Coarse-grained energy models are based on statistical preferences of residues or atoms to adopt distinct states. They can be expressed as an energetic term according to the inverse Boltzmann's law and are also referred to as potentials of mean force [192, 193, 191, 194, 195]. A common approach is e.g. the distinction of residues which are exposed to the solvent and those which do not interact with the solvent at all, because these residues are buried in the hydrophobic core of a protein [183]. Intuitively, these propensities change between amino acids. Hydrophobic ones such as phenylalanine or methionine prefer isolation from the polar solvent, whereas hydrophilic amino acids such as arginine or glutamate can energetically favorably interact with surrounding water molecules but not with hydrophobic amino acids or lipid molecules of the cell membrane. The derived values are not directly related to physical energy values; however, their applicability has been proven in problems

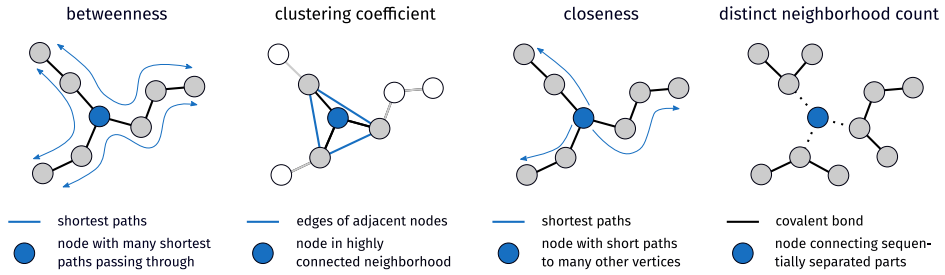

Figure 5.2.: Topological Properties Used for the Characterization of Early Folding Residues

The betweenness describes how many shortest paths pass through a node [189, 186]. The closeness is the inverse of the average path length to all other nodes [190]. The clustering coefficient states how well-connected adjacent nodes of a node are [186]. The distinct neighborhood count captures how many sequentially separated regions are connected by a residue.

of molecular docking [196], threading and fold recognition [197, 191], homology modeling [194], and model quality assessment programs [194, 198, 182].

There are various ways of implementing and parameterizing such coarse-grained energy models [193, 199, 198] with their differences boiling down to the exact definition of what constitutes a contact between residues. All approaches share the fact that they are merely models of reality and all have advantages and drawbacks. The approach of so-called Energy Profiles [183] is another coarse-grained model which the major advantage that a full-fledged ecosystem of methods for the computation from structure, prediction from sequence (by the eGOR method), and the alignment of Energy Profiles exists. All this functionality is provided by the energy profile suite (eProS) [183].

Energy Profiles were calculated from structure and predicted from sequence according to the methodology [183] used in the eQuant web server. For each amino acid, the respective propensity of it to occur exposed to the solvent or buried in the hydrophobic core was determined. A residue i is categorized as buried if:

$$||C_{\alpha,i} - c|| < 5\ \text{Å} \vee (C_{\alpha,i} - C_{\beta,i})(C_{\alpha,i} - c) < 0 \tag{5.1}$$

Thereby, c refers to the centroid of all C_α atoms less than 5 Å away from $C_{\alpha,i}$. Using this criterion, the absolute counts for each amino acid to occur exposed ($n_{aa,exposed}$) or buried ($n_{aa,buried}$) are determined. These propensities can be expressed as pseudo-energies according to the inverse Boltzmann's law:

$$e_i = -k_B T \cdot ln \frac{n_{aa,buried}}{n_{aa,exposed}} \tag{5.2}$$

The energy of a residue i is defined as the sum of these pseudo-energies for all pairs of interacting residues:

$$E_i = \sum_{j \in S \setminus i} g(i,j)(e_i + e_j) \tag{5.3}$$

Interactions are defined by the function $g(i,j)$ as residues for which the distance of the C_β atoms is less than 8 Å – the position of the C_α atom is used as fallback for glycine. In summary, Energy Profiles describe each residue by an energy value which captures the surrounding of each residue.

The eQuant web server takes protein structure in PDB format as input and computes the likely structural uncertainty of each residue (i.e. local quality assessment in Figure 5.3). Several structural features are computed to calculate features describing each residue. Subsequently the local error of each residue is calculated by the random subspace method [200]

implemented in Waikato environment for knowledge analysis (Weka) [201, 202, 203]. This information is then used to compute the global quality score of the entire protein which is related to scores of proteins of similar size for comparability.

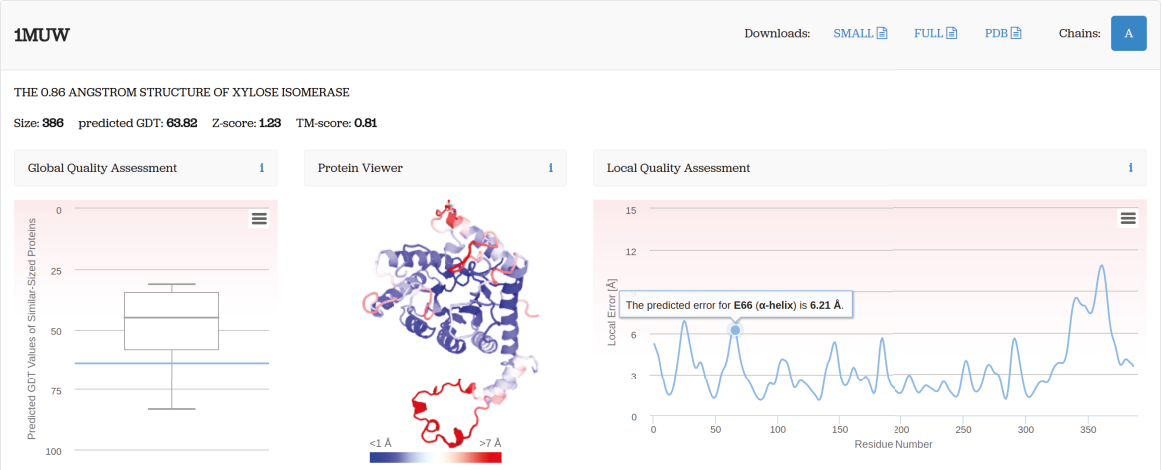

Figure 5.3.: User Interface of the eQuant Web Server

eQuant analyzes the quality of protein structures and *in silico* models. For the global quality assessment (left panel), the overall score is compared to proteins of similar size. The structure is furthermore colored by the local quality scores (central panel). The computed local per-residue error values are plotted as line chart (right panel). All plots are interactive: hovering over certain positions in the line chart will highlight the appropriate residues in the structure and vice versa.

5.1.4. FEATURE COMPUTATION

RASA values were computed by the algorithm of Shrake and Rupley [166]. Buried residues are defined as those with RASA values less than 0.16 [168]. Non-covalent residue-residue contacts were detected by PLIP [37]. Secondary structure elements were annotated using DSSP [167]. For both ASA and secondary structure element annotation the BioJava [164, 165] implementations were used. The loop fraction is defined as fraction of unordered secondary structure in a window of nine residues around the evaluated amino acid [204]. This yields a fraction, where high values are tied to regions of high disorder, whereas amino acids embedded in α-helices or β-sheets result in scores close to zero. The centroid distance of a residue is the spatial distance of its centroid to that of all atoms. The terminus distance is the minimal value of sequence positions to either terminus divided by the number of residues. Evolutionary information as well as evolutionary co-variation scores were computed using the EVfold web server [140, 141]. The evolutionary information is based on the MSA of homologues automatically retrieved for the query sequence and expresses how conserved a column in this MSA is.

5.1.5. STATISTICAL ANALYSIS

The Start2Fold database [8] was utilized to assess whether the separation EFR and functional residues is a common theme in protein structures.

Various features were used to describe residues of the dataset and it was tested whether these features differ significantly between EFR and LFR as well as EFR and functional residues. In both cases, *p*-values were computed on the subset of buried residues (RASA less than 0.16 [168]), because EFR tend to be buried in the hydrophobic core of proteins [13] whereas functional residues are likely exposed to the solvent. All tables present the average

and standard deviation of the considered features for all residues and the corresponding *p*-value for the subset of buried residues. Also, the *p*-value for buried residues is used when the level of significance is stated. Dependence of distributions of real-valued variables was tested by the Mann-Whitney U test. Dependence of distributions of count variables was tested using the Dunn test with Bonferroni correction. Throughout the manuscript, * corresponds to significant *p*-values <0.05 for the Mann-Whitney U and *p*-values <0.025 for the Dunn test. A variation of conventional boxplots is used, which depict a notch around the median of the distribution. This notch corresponds to the 95% confidence interval and allows to visually assess whether the medians of two boxes are similar. No overlap of notches indicates that both medians differ substantially [205].

5.2. RESULTS & DISCUSSION

A previously described dataset [8] of 27 proteins and 2,966 residues is the basis of this chapter and summarized in Table 5.1. 450 (15.2%) of the residues are labeled as EFR, the remaining residues are considered LFR. Hydrophobic amino acids have been previously described to have a higher propensity of being EFR [13].

To characterize EFR in more detail, various features were defined and compared to the values of LFR. EFR form a significantly greater number of residue-residue contacts (i.e. distance less than 8 Å) than their LFR counterparts (Figure 5.4A). The loop fraction is defined as the ratio of unordered secondary structure elements in a window centered on a particular residue [204]. Fewer unordered secondary structure elements can be found around EFR (Figure 5.4B), whereas LFR exhibit a higher propensity to occur in coil regions. EFR are on average closer to the centroid of a protein structure and are likely embedded in the hydrophobic core (Figure 5.4C). Analogously, they also tend to be more distant to the N- or C-terminus of the sequence than other residues and are likely buried regarding their RASA as per Table 5.2.

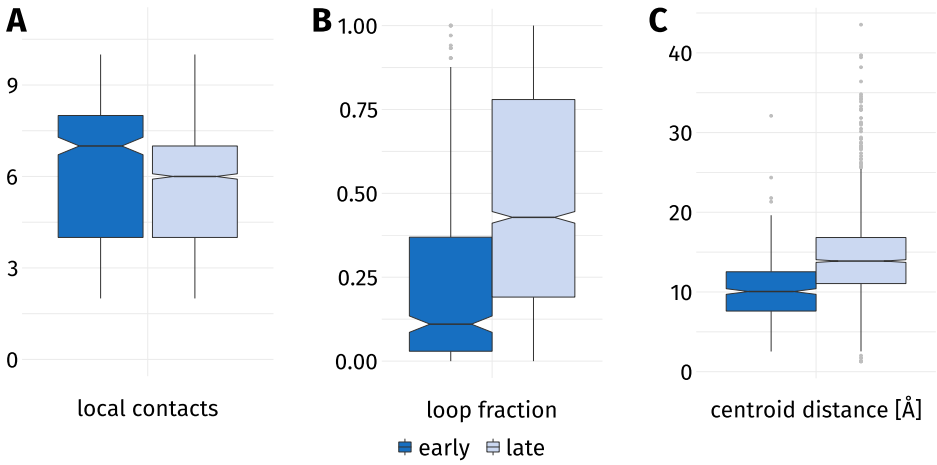

Figure 5.4.: General Properties of Early and Late Folding Residues

(**A**) EFR (dark blue) form more contacts to their surroundings than LFR (light blue). (**B**) The loop fraction [204] is the ratio of unordered secondary structure elements which are observed in a windows of nine amino acids around a residue. EFR are more commonly surrounded by ordered secondary structure elements. (**C**) EFR are located significantly closer to the centroid of the protein than LFR.

The propensity of EFR to participate in more contacts and to occur in the core of a protein are in agreement with previous studies [13, 18, 186, 20]. The shift in loop fraction can also be attributed to these findings and is further substantiated by the fact that long ordered

Table 5.2.: Statistical Characterization of Early Folding Residues

feature	μ_{early}	σ_{early}	μ_{late}	σ_{late}	p_{buried}	level
# contacts	11.12	2.46	9.02	2.92	0.0004	*
loop fraction	0.22	0.26	0.48	0.33	<0.0001	*
centroid distance [A]	10.28	3.79	14.14	4.81	<0.0001	*
terminus distance	0.27	0.14	0.25	0.15	0.0035	*
RASA	0.16	0.18	0.34	0.26	<0.0001	*
computed energy	-17.57	11.63	-9.40	9.86	<0.0001	*
predicted energy	-16.89	10.78	-11.07	10.31	<0.0001	*
betweenness	0.06	0.03	0.04	0.03	<0.0001	*
closeness	0.32	0.06	0.28	0.06	<0.0001	*
clustering coefficient	0.51	0.10	0.59	0.14	0.5688	-
# distinct neighborhood	2.47	1.53	1.88	1.54	0.0446	*
# local contacts	6.28	1.84	5.82	1.65	0.0532	-
# tertiary contacts	4.85	3.17	3.20	2.96	0.0575	-
# hydrogen bonds	3.85	1.43	2.84	1.79	0.0001	*
# hydrophobic interactions	1.31	1.46	0.65	1.07	<0.0001	*
# coupling	2.8	1.89	1.98	1.88	0.0903	-
cumulative strength	0.72	0.60	0.51	0.57	0.7638	-
coupling strength	2.93	2.13	1.96	2.00	0.0522	-
evolutionary information	40.47	26.01	28.34	26.46	0.0122	*

For each presented feature the mean (μ) and standard deviation (σ) of both the EFR and LFR category is reported. It was tested whether the differences of a feature between EFR and LFR state are significant. p_{buried} refers to the p-value of the test on residues buried according their RASA value, this was done because EFR have a tendency to be located in the core of a protein and without filtering all differences are significant. The Mann-Whitney U test was used for real-valued variables, whereas the Dunn test was used for count variables (indicated by #). 2,966 residues in 27 proteins from the Start2Fold database [8] were analyzed. Features and employed tests are described in Section 5.1.

secondary structure elements tend to contain more EFR [13]. It has been reported that buried residues are more likely to be EFR [87, 8] which also explains why they are closer to the spatial centroid of a protein and more separated from sequence termini (Table 5.2). Evolutionary couplings scores reported by the DCA [140, 141] and evolutionary information exhibit interesting properties: the reported coupling strength as well as evolutionary information of EFR are significantly increased. The relation of evolutionary information and EFR has been the subject of previous studies [13, 11]. All these factors can neither explain why some residues become EFR while others do not nor how EFR relate to the rest of a protein in terms of network analysis.

The computed energy of EFR is significantly lower than the values of LFR.Regarding the average absolute contact frequencies, a EFR participates in 3.85 hydrogen bonds and forms 1.31 hydrophobic interactions with other residues. This constitutes a significant increase compared to LFR (see Table 5.2).

The low computed energies indicate that EFR have an intrinsic propensity to form stable, local conformations. EFR might be the mediators between the formation of local structure elements and their assembly in the context of the three-dimensional structure. Secondary structure elements such as helices interact e.g. by hydrophobic interactions [206], however, it seems that single contacts are neither strong nor specific enough to guide their assembly [207, 33, 19]. A fine-grained distinction of contact types including π-stacking and hydrophobic interactions would be required to assess the role of EFR as potential driving

force behind the correct of arrangement of secondary structure elements.

To gain additional insights, a second population of residues relevant for the folding process was analyzed (Table 5.3): HSR are residues with superior stability regarding to unfolding events. These residues may not be relevant for the initial formation of the native structure, but seem to prevent spontaneous unfolding [161]. Therefore they are a different set of residues which may be structurally relevant. HSR again show a bias to be buried residues and assessment of statistical significance leads to similar results as for EFR. Interestingly, HSR feature an increased number of evolutionary couplings and also show a significant increase in tertiary contacts (i.e. a characteristic not observed to differ significantly for EFR). Overall, the characteristics of EFR and HSR are comparable.

Table 5.3.: Statistical Characterization of Highly Stable Residues

feature	μ_{stable}	σ_{stable}	$\mu_{unstable}$	$\sigma_{unstable}$	p_{buried}	level
# contacts	11.16	2.42	8.84	2.89	<0.0001	*
loop fraction	0.25	0.27	0.50	0.33	<0.0001	*
centroid distance [Å]	10.45	4.09	14.40	4.72	<0.0001	*
terminus distance	0.26	0.14	0.25	0.14	0.0909	-
RASA	0.17	0.19	0.36	0.26	<0.0001	*
computed energy	-17.71	11.89	-8.69	9.27	<0.0001	*
predicted energy	-16.85	10.92	-10.61	10.09	<0.0001	*
betweenness	0.06	0.03	0.04	0.03	<0.0001	*
closeness	0.32	0.07	0.28	0.06	<0.0001	*
clustering coefficient	0.51	0.10	0.59	0.15	0.6468	-
# distinct neighborhood count	2.51	1.52	1.82	1.53	0.0026	*
# local contacts	6.24	1.85	5.79	1.63	0.3680	-
# tertiary contacts	4.92	3.20	3.05	2.88	0.0030	*
# hydrogen bonds	3.79	1.41	2.77	1.81	<0.0001	*
# hydrophobic interactions	1.25	1.46	0.62	1.02	<0.0001	*
# coupling	2.92	2.01	1.88	1.81	<0.0001	*
cumulative strength	0.77	0.64	0.48	0.55	<0.0001	*
coupling strength	3.05	2.24	1.85	1.91	<0.0001	*
evolutionary information	40.23	26.80	27.43	26.07	0.0056	*

For each presented feature the mean (μ) and standard deviation (σ) of both the HSR and UR category is reported. p_{buried} refers to the p-value of the test on residues buried according their RASA value. This was done because HSR have a tendency to be located in the core of a protein and without filtering most differences are significant. Features and employed tests are described in Section 5.1.

5.2.1. NETWORK ANALYSIS SHOWS A UNIQUE WIRING OF EARLY FOLDING RESIDUES

The way residues interact with their spatial surrounding was assessed by network analysis based on residue graphs. Regarding the topological properties of residues derived from network analysis (see Figure 5.2 for a graphical depiction), EFR are more connected than LFR. They exhibit higher betweenness (Figure 5.5A) and closeness (Figure 5.5B) values. High betweenness values are observed for well-connected nodes which are passed by many of shortest paths in a graph. High closeness values occur for nodes which can be reached by relatively short paths from arbitrary nodes. The distinct neighborhood count expresses how many sequentially separated regions of a protein a residue is connected to. Again a significant increase can be observed for EFR (Figure 5.5C). Residues are considered

separated when they are more than five sequence positions apart. This threshold was also used to distinguish local contacts (i.e. less than six residues apart) and tertiary contacts. Interestingly, the clustering coefficient features a significant decrease when EFR are considered. The clustering coefficient of a node is the number of edges between its adjacent nodes divided by the theoretical maximum of edges these nodes could form. However, EFR are biased to be in the core of the protein [13], thus, it was assessed if this change is also significant when only buried [168] residues are considered. The differences regarding the clustering coefficient are insignificant in that case (see Table 5.2).

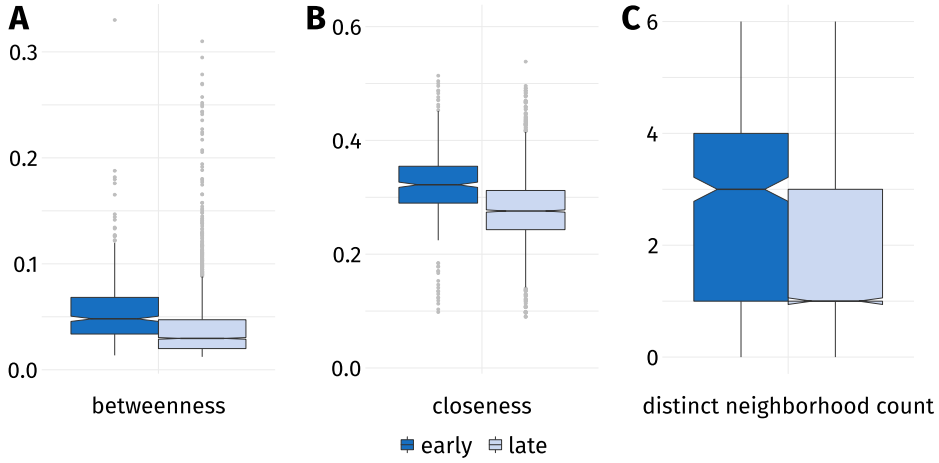

Figure 5.5.: Topological Properties of Early and Late Folding Residues

Proteins were represented as residue graphs and network analysis was performed. (**A**) EFR have higher betweenness values implying that shortest paths in the graph tend to pass through these nodes more often. (**B**) They also exhibit higher closeness values because their average path length to other nodes is lower. (**C**) The distinct neighborhood count of a residue describes to how many separated regions it is connected. Residues are considered separated when their separation at sequence level is greater than five. EFR connect significantly more regions of a protein than LFR.

The betweenness property is closely related to the small-world characteristics of networks (i.e. they are well-connected even when between most nodes no edge is present) and can be observed in this case due to the ratio of protein surface and volume [186]. Residues relevant for the folding process have been shown to exhibit high betweenness values in the transition state and to be crucial for the formation of the folding nucleus [186]. Interestingly, the clustering coefficient shows no difference between EFR and LFR when only buried residues are considered. Also, the value is higher for LFR, which is probably an effect of EFR being hubs which connect several separated regions of a protein (as shown by the distinct neighborhood count). These regions themselves are not well-connected, which results in a lower clustering coefficient for EFR. The performed network analysis aids the understanding on the idiosyncratic properties of EFR in the context of the whole protein and is in agreement with previous studies [186, 94, 208]. EFR are hubs between sequentially distant protein regions which underlines their importance for the correct assembly of the tertiary structure of a protein. Nevertheless, the increased number of local and tertiary contacts of EFR point to their importance for the whole protein folding process as described by the defined-pathway model [18, 19]. The protein folding process is difficult to study due to various aspects such the existence of disordered proteins [209, 20], the relevance of chaperons [209], cotranslational folding [86], and the insertion of membrane proteins by the translocon [206, 210]. EFR are a welcome simplification to advance the understanding.

5.2.2. EARLY FOLDING AND FUNCTIONAL RESIDUES EXHIBIT DISTINCT FEATURES

Division of labor is one of the most successful strategies of evolution [211, 24, 23, 22]. The separation of residues crucial for folding and those furnishing function may allow reuse of established protein folds [24, 212, 23, 22]. The sequence and structure space ascertained over the course of evolutions seems small for a truly random exploration. Reusing established folds could also avoid slow-folding sequences or those prone to aggregation [102, 212, 213]. There seems to be a delicate balance in proteins between robustness and evolvability [24, 214, 55]. Functional residues [25] can be mutated and new functions can be adapted without compromising the fold of the protein [111]. In consequence, a clear division should be observable between EFR – which initiate and guide the folding process – and the functional ones implementing protein function.

To address this question, residues in the dataset were labeled as either EFR or LFR as well as either functional or non-functional. Active sites and ligand binding regions were considered to be the functional parts of proteins. The distribution of both binary variables (Table 5.4) shows that the majority of residues in the dataset are neither EFR (86.1%) nor functional (93.9%) residues. 0.9% share both classes, whereas 0.8% are expected to share both classes if their association was random (see Section 5.1.5). The distribution of both variables separated by individual proteins is presented in Table 5.1. For many proteins, no residues are both EFR and functional (Figure 5.6A). Furthermore, EFR tend to be located in the core of proteins, whereas functional residues are exposed towards the solvent in order to realize their respective function (Figure 5.6). The acyl-coenzyme A binding protein (STF0001) [215, 216] features five residues which are both EFR and functional (Figure 5.6B). Another case where the overlap is large is T-cell surface antigen CD2 (STF0009) which can bind other protein antigens.

Table 5.4.: Contingency Table of Early Folding Characteristics and Functional Relevance

	functional	non-functional
early	22	324
late	130	2,014

Out of 2,490 observations, 0.9% are EFR and functional at the same time. Based on the presented frequencies, 0.8% of all residues are expected to share both labels if their association is independent. This captures that a separation of EFR and functional cannot be observed in general. Proteins were excluded when no annotation of functional residues existed.

The acyl-coenzyme A binding protein may exhibit five residues which are both EFR and functional because its a rather small protein of 86 residues which binds ligands with large aliphatic regions. Intuitively, the residues furnishing the bowl-like shape of the protein are also those which participate in the function of ligand binding [215, 216]. Roughly half the residues of the acyl-coenzyme A binding protein are marked as EFR which further accentuates why no separation can be observed in this case. Exceptionally well-separated are EFR and functional residues in the fibroblast growth factor 1 (STF0024) and Villin-1 (STF0028). The first protein contains a large number of EFR distributed throughout the sequence and a large functional heparin-binding region which are distinct at sequence level. Villin-1 exhibits a similar distribution of EFR and features a C-terminal polyphosphoinositide binding region which contains no EFR. In both cases, the functional sites bind other molecules. This characteristic is commonly associated to increased structural flexibility [217] which may explain why EFR rarely occur there. The primary selection pressure during evolution is on protein function [4] rather than on structural integrity [218]. In cases where a certain position is crucial for function, slower folding is tolerated which implies that structure and folding are

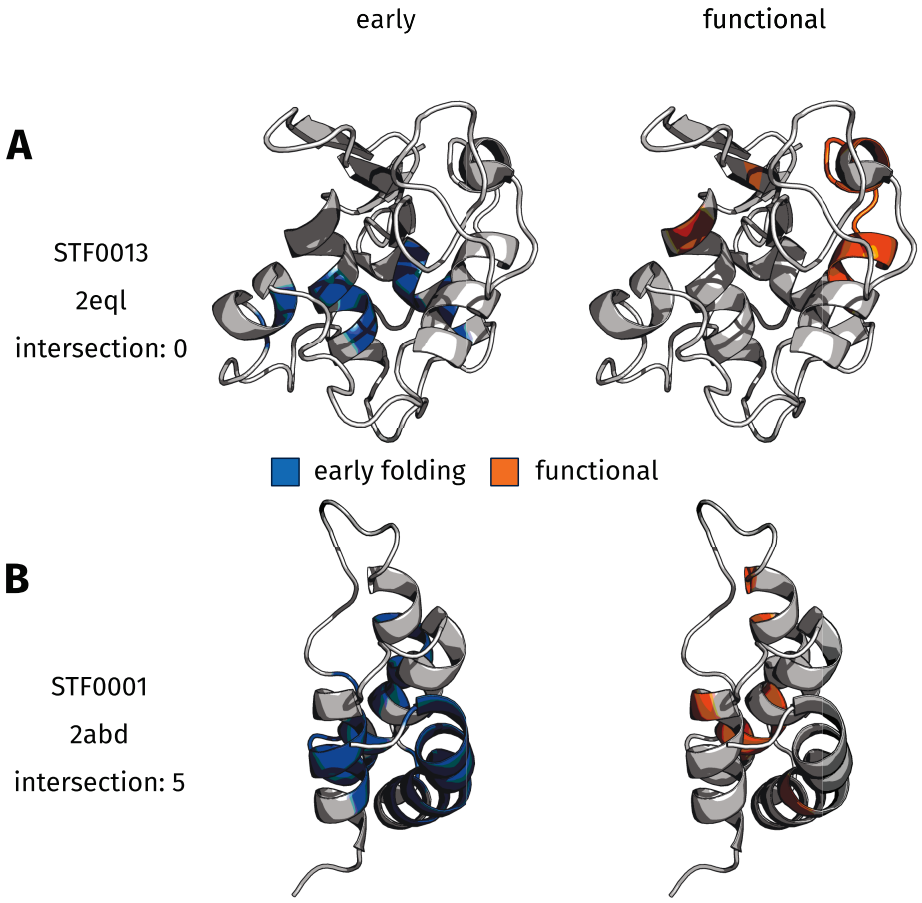

Figure 5.6.: Rendered Structures of Two Dataset Entries

EFR are rendered in blue, functional residues are rendered in orange. (**A**) In the case of lysozyme (PDB:2eql_A) the intersection of EFR and functional residues is empty. For most proteins in the dataset, there is a clear distinction between both classes and structurally relevant residues have a propensity to be located in the core, while functional residues are exposed on the surface of the protein. (**B**) Five residues are both EFR and functional in the acyl-coenzyme A binding protein (PDB:2abd_A) which is one of the exceptions in the dataset where some residues are both EFR as well as functional.

subordinated to function [55]. Disordered proteins are another example of proteins without structural integrity which achieve a high robustness of function [188]. In structural biology, structure is considered to be a scaffold which allows proteins to implement a particular function [188, 4]. During evolution, it is most important that proteins retain their function [219, 4] and this may even require an explicit lack of a defined structure or structural flexibility [217]. This potential irrelevance of a particular fold underlines that the separation of structurally and functionally relevant residues may be advantageous for evolvability. However, cases have been described where it is advantageous to place functional residues close to residues ensuring structural integrity in order to maintain protein function over the course of evolution [29]. Another interpretation with respect to the defined-pathway model [18] is that EFR initiate and guide the folding process – though other residues may be of higher structural relevance in the native structure. Nevertheless, by assigning this responsibility to a small number of residues, the remaining residues can constitute active sites.

Interestingly, the separation of structurally relevant residues and those implementing protein function can also be observed for HSR (Table 5.5). Therein, the expected frequency of residues sharing both the functional label and that of HSR is higher than the overlap actually observed. This substantiates the previous finding from a different point of view.

Table 5.5.: Contingency Table of Highly Stable Characteristics and Functional Relevance

	functional	non-functional
stable	28	510
unstable	124	1,828

Out of 2,490 observations, 1.1% are HSR and functional at the same time. Based on the presented frequencies, 1.3% of all residues are expected to share both labels if their association was independent. This captures the tendency to separate HSR and functional residues. Proteins were excluded when no annotation of functional residues existed.

The previously described features were employed to substantiate the identified separation of structure and function at residue level (Table 5.6). EFR show significantly lower computed energies when compared to LFR or functional residues (Figure 5.7A). Functional residues exhibit higher computed energies than their non-functional counterparts. Most residues form only a small number of hydrophobic interactions, however, the number is significantly increased for EFR (Figure 5.7B). 97.6% of EFR form hydrogen bonds and 65.1% participate in hydrophobic interactions. Functional residues participate to 88.8% in hydrogen bonds and to 39.5% in hydrophobic interactions. On the contrary, the change between the hydrogen bond count of EFR and functional residues in a buried state is insignificant (Table 5.6). The clustering coefficient of a node captures how many edges can be observed between the adjacent nodes and, thus, describes how well-connected the direct surroundings of a node are. Functional residues show an insignificant change regarding this property (Table 5.6). In contrast, the clustering coefficient significantly decreases when EFR are compared to LFR or functional residues (Figure 5.7C). In summary, EFR exhibit distinct properties compared to functional residues. Their surrounding secondary structure elements, computed energy values, and the number of hydrophobic interactions are especially characteristic. In terms of evolutionary information, functional residues exhibit a significant change compared to non-functional residues (Table 5.6). When buried, evolutionary information of functional residues amounts to 43.39 compared to 42.40 for EFR. LFR and non-functional residues are less conserved at sequence level.

Due to their purpose, EFR are located in the hydrophobic core and functional residues are primarily exposed to the solvent. These distinct requirements manifest in the computed energies. Furthermore, protein function can commonly be broken down to amino acids which feature hydrophilic, chemically functional groups [25]. Hydroxyl groups are a prominent examples for functional groups contributing to catalysis [25]. Thus, functional residues are likely to exhibit above average computed energies because of their higher propensity to contain hydrophilic side chains. Analogously, fewer hydrophobic amino acids constitute the functional residues of binding sites and they form fewer hydrophobic interactions. Most of the hydrophobic interactions are accumulated in the hydrophobic core of a protein [27, 183, 220]. EFR tend to be crucial connectors in proteins, however, their clustering coefficient is low. This can be attributed to the fact that EFR connect many distinct neighborhoods. Furthermore, functional residues feature above average closeness values: they are well-connected to other parts of the protein, even though they are unaffected by the early folding events. It was shown that functional residues have special requirements on how they are wired to the rest of a protein [190]: surrounding residues ensure the correct placement of functional residues [221, 222, 190], modulate their chemical properties such as pK_a values [25, 223, 190], or propagate signals to other parts of a protein [190]. Analogously, the evolutionary pressure on functional residues is increased compared to EFR and non-functional residues as indicated by the evolutionary information (Table 5.2). In particular, catalytic activity of amino acids can be broken down to functional groups of their side chain [25]. The hydroxyl side chain of serine may be substituted by threonine

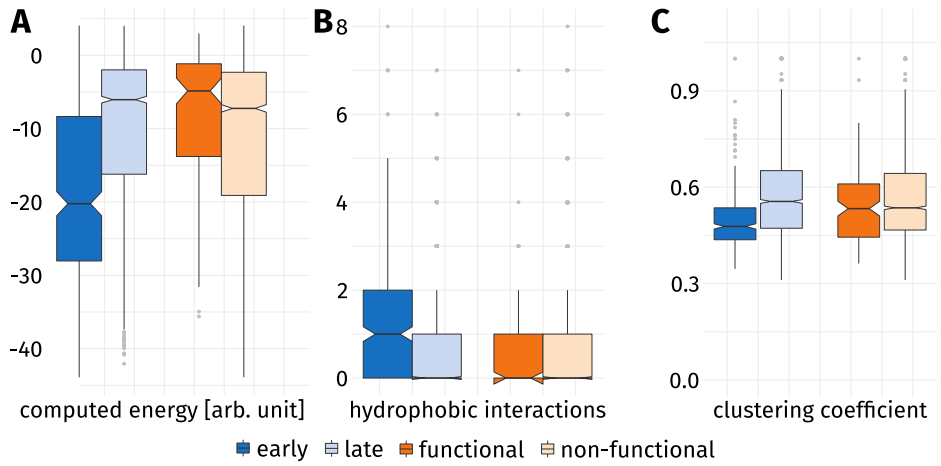

Figure 5.7.: Characteristics of Early Folding and Functional Residues

EFR (dark blue) and LFR (light blue) are compared to functional (dark orange) and non-functional (light orange) residues. (**A**) EFR show lower computed energies than they are in contact with many residues and tend to be embedded in the hydrophobic core. In contrast, functional residues are exposed to the solvent in order to constitute e.g. binding sites. (**B**) Hydrophobic interactions occur especially in the core of a protein, thus, most residues do not form any. However, EFR show a significant increase compared to LFR. (**C**) The clustering coefficient of a node describes how well-connected its adjacent nodes are. EFR connect regions of a protein which are separated at sequence level and, thus, not well-connected on their own. Functional residues exhibit higher clustering coefficient indicating a more connected set of adjacent nodes.

Table 5.6.: Comparison of Early Folding and Functional Residues

feature	$p_{\text{func/non,buried}}$	μ_{early}	μ_{func}	$p_{\text{early/func,buried}}$	level
# contacts	0.1268	11.31	9.39	0.0165	*
loop fraction	0.9057	0.24	0.43	0.0089	*
centroid distance [Å]	0.0570	10.68	12.13	0.9370	-
terminus distance	0.0503	0.27	0.25	0.0072	*
RASA	0.0800	0.15	0.30	0.0034	*
computed energy	<0.0001	-17.99	-8.18	<0.0001	*
predicted energy	0.0028	-17.15	-10.23	<0.0001	*
betweenness	0.1441	0.05	0.05	0.6497	-
closeness	0.4087	0.31	0.27	0.3282	-
clustering coefficient	0.1435	0.51	0.58	0.2328	-
# distinct neighborhood count	0.1753	2.52	2.24	0.3666	-
# local contacts	0.2177	6.18	5.70	0.1457	-
# tertiary contacts	0.4084	5.13	3.69	0.1558	-
# hydrogen bonds	0.0911	3.84	3.17	0.4975	-
# hydrophobic interactions	0.0019	1.29	0.60	0.0001	*
# coupling	0.3601	2.85	2.51	0.4996	-
cumulative strength	0.3792	0.74	0.63	0.4086	-
coupling strength	0.7575	2.97	2.29	0.3219	-
evolutionary information	0.0021	42.40	43.39	0.0299	*

For each presented feature the distribution of values is compared between functional and non-functional residues as well as EFR and functional residues. The corresponding *p*-values and significance level are stated for buried residues. Mean values are shown for EFR (μ_{early}) and functional residues (μ_{func}). Features and employed tests are described in Section 5.1.

or tyrosine. In contrast, contacts which stabilize protein structures can be primarily broken down to the hydrophobic or hydrophilic character of amino acids [42, 43] which allows for a wider range of tolerated mutations. Early stages of protein folding sample transient conformations [18, 13] and settle for stable, local structures as indicated e.g. by the Energy Profiling approach. It has been shown that the characteristic of EFR is not directly linked to individual amino acids but rather the effect of the sequence composition of larger fragments [20, 13, 11]. This may be another explanation why EFR are less conserved at sequence level than functional residues. That the folding nucleus of proteins is not necessarily sequentially conserved has been demonstrated previously [224, 225, 18], and makes it even more remarkable that coevolution techniques such as the DCA perform so well for structure prediction tasks [140, 141].

Modularity in proteins is also present in domains [23], secondary structure elements, and autonomous folding units of the defined-pathway model [19, 82]. Particularized knowledge of EFR may improve synthetic biology and could allow the design of proteins combining existing functional domains without influencing one another negatively [226, 24, 28, 23]. Furthermore, understanding the differences of structurally relevant residues and those implementing function could help in predicting mutation effects and provide a new level of detail by allowing whether a mutation disrupts the fold or the function of a protein [227, 144].

5.3. CONCLUSION

A dataset of EFR for the protein folding process was studied. They are highly connected nodes in residue graphs and were observed to be located in energetically favorable conformations as pointed out by the Energy Profiling approach [183]. These structurally relevant residues have distinct properties e.g. regarding the number of hydrophobic interactions compared to functional residues.

Future HDX data can substantiate the presented trends regarding the nature of EFR. Potentially, the arsenal of experimental techniques to study the folding process of proteins will expand and become more refined and standardized, so that the underlying dataset will become more robust. Also the observed separation of structurally relevant and functional residues in proteins may be substantiated by more data on EFR and HSR. Understanding these topological differences provides insights into the way certain residues interact with the rest of the protein and to what degree they tolerate or compensate manipulation. For decades, scientists longed for a glimpse into the folding process [9, 10, 11] and the analyzed dataset [8] provides just that. The experimental signals of early folding events are still difficult to interpret and the analyzed dataset may not be generalizable for large proteins, but the made observations indicate that EFR are also relevant as structural hubs in the native structure.

EFR are a valuable to gain insights into the folding process with spatial and temporal resolution. Future studies may link them to characteristics at sequence level to understand the sequence composition which causes particular regions of a protein to initiate the folding process. Features presented in this study were shown to be discriminative for EFR. Classifiers for them based on sequence [11] or structure data may annotate residues crucial for protein folding. Trained classifiers can also report as well as visualize the most discriminative features [228, 229] which may further delineate EFR and LFR. The Chapter 6 presents a machine learning strategy which is employed with this aim on the Start2Fold dataset [8].

6. AN INTERPRETABLE CLASSIFICATION MODEL FOR EARLY FOLDING RESIDUES

This Chapter is Based on the Publication:

Bittrich, S., Kaden, M., Leberecht, C., Kaiser, F., Villmann, T., & Labudde, D. (2018). Application of an Interpretable Classification Model on Early Folding Residues during Protein Folding. *BioData Mining*, in press.

This Chapter Employs Methods from:

Bittrich, S., Heinke, F., & Labudde, D. (2016). eQuant – A Server for Fast Protein Model Quality Assessment by Integrating High-Dimensional Data and Machine Learning. In *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery* (pp. 419-433). Springer, Cham.

The analysis of data collected during biological experiments poses a challenge for modern bioinformatics. Usually these data are feature rich, yet hard to interpret, such as single-cell gene expression data obtained by high-throughput experiments [230]. Despite sophisticated pre-processing and the application of machine learning models, analysis – and most importantly interpretation – of such data is still hard to accomplish. Nevertheless, machine learning is the basis for sophisticated predictions and allows new insights into open questions. An interpretable classifier on the protein folding problem was developed in order to deepen the understanding of EFR based on the trained model.

This Chapter focuses on interpretability and discussion of the resulting model. It is demonstrated how an adaptation of an established machine learning strategy allows pinpointing the most influential features for classification. Therefore, a novel implementation of the generalized matrix learning vector quantization (GMLVQ) algorithm [231, 232] as plug-in for the popular Weka framework [201, 202, 203] is presented. This plug-in features diverse visualization tools which encourage the user to interpret the resulting model and render GMLVQ a comprehensible *white box* classifier.

6.1. MATERIALS & METHODS

6.1.1. DATASET CREATION

The Start2Fold dataset [8] and features presented in Chapter 5 are the basis of this analysis. All 111 proline instances were dropped from the initial dataset which resulted in 3,266 residues of which 482 (14.8%) are EFR. All considered features describe a particularized aspect of this connection and are summarized in Table 6.1. Features of each residue were averaged with respect to four adjacent positions at the sequence level in N- and C-terminal direction. The dataset was standardized by z-score transformation.

6.1.2. FEATURE COMPUTATION

For a more detailed description of individual features refer to Chapter 5.

Energy Profiling Energy Profiles [183, 182] transform the three-dimensional arrangement of atoms in a protein into a vector of energy values describing each amino acid. The computed energy (*e*) of a residue describes its interactions with its surroundings. Energy Profiles can also be predicted using only sequence information [183] (*ePred*) which represents the sequence composition. Computed as well as predicted energy values have been used before for the description of the folding process [183] as well as protein structure quality assessment [182].

Secondary Structure Elements Secondary structure elements were annotated using DSSP [167]. The secondary structure element size of a residue (*SecSize*) refers to the number of sequence neighbors sharing its secondary structure (i.e. α-helix, β-strand, and coil). For sequence windows of nine residues the number of unordered secondary structure elements was counted and normalized by the window size [204]. This yields a fraction (*LF*), where high values are tied to regions of high disorder, whereas amino acids embedded in α-helices or β-sheets result in scores close to 0.

Relative Accessible Surface Area The RASA of a residue describes how exposed it is to the solvent. Residues in the hydrophobic core tend to be buried and exhibit no accessible surface area. RASA values (*Rasa*) were computed with the algorithm of Shrake and Rupley [166].

Table 6.1.: Denomination and Short Description of the 27 Features of the Dataset

feature	description
e	computed energy values
ePred	predicted energy values
SecSize	size of the surrounding secondary structure elements
LF	fraction of surrounding unordered secondary structure elements
Rasa	relative accessible surface area
PlipLC	absolute count of local PLIP contacts
PlipHbLC	absolute count of local PLIP hydrogen bonds
PlipHpLC	absolute count of local PLIP hydrophobic interactions
PlipBbLC	absolute count of local PLIP backbone contacts
PlipLR	absolute count of tertiary PLIP contacts
PlipHbLR	absolute count of tertiary PLIP hydrogen bonds
PlipHpLR	absolute count of tertiary PLIP hydrophobic interactions
PlipBbLR	absolute count of tertiary PLIP backbone contacts
PlipBN	betweenness using all PLIP contacts
PlipCL	closeness using all PLIP contacts
PlipCC	clustering coefficient using all PLIP contacts
PlipHbBN	betweenness using PLIP hydrogen bonds
PlipHbCL	closeness using PLIP hydrogen bonds
PlipHbCC	clustering coefficient using PLIP hydrogen bonds
PlipHpBN	betweenness using PLIP hydrophobic interactions
PlipHpCL	closeness using PLIP hydrophobic interactions
PlipHpCC	clustering coefficient using PLIP hydrophobic interactions
ConvBN	betweenness using the distance-based contact definition
ConvCL	closeness using the distance-based contact definition
ConvCC	clustering coefficient using the distance-based contact definition
PlipNC	distinct neighborhood count using all PLIP contacts
ConvNC	distinct neighborhood count using the distance-based contact definition

References to these features are given in *italic* font.

Non-Covalent Contacts Non-covalent contacts stabilize protein structures and are the driving force behind protein folding [233]. The PLIP [37] was used for the annotation of non-covalent contacts between residues in protein structures. PLIP supports different contact types such as salt bridges, π-stacking interactions, or π-cation interactions. For this analysis, only hydrogen bonds (*Hb*) and hydrophobic interactions (*Hp*) were considered. Other contact types were not observed for the rather small proteins in the dataset. Furthermore, local and tertiary contacts [123] were distinguished. Local contacts (suffix *LC*) are defined as contacts between residues less than six sequence positions apart – their main contribution is stabilizing secondary structure elements. In contrast, tertiary contacts (suffix *LR*) occur between residues more than five sequence positions apart and constitute stabilizing contacts between secondary structure elements which primarily manifest the three-dimensional arrangement of a protein. Backbone contacts (*Bb*) occur only between backbone atoms of the respective residues.

Graph Representation of Proteins Proteins in the dataset were represented as graphs. Amino acids always constituted the nodes and contacts between residues were represented by edges. Covalently bound residues were considered to be in contact. All contacts annotated by PLIP were used to create the first graph representation (using the prefix *Plip*). Reduced representations were created by only considering hydrogen bonds (using the prefix *PlipHb*) respectively hydrophobic interactions (using the prefix *PlipHp*). The contacts detected by PLIP may ignore spatially close residues when they do not form any contacts according the underlying rule set. Therefore, an additional contact definition was employed (prefix *Conv*): two residues were considered to be in contact, if their C_α atoms were at most 8 Å apart.

Topological Descriptors Based on the four graph representations (*Plip*, *PlipHb*, *PlipHp*, and *Conv*), topological descriptors of individual residues were computed. This allows to describe how residues are connected to other residues by means of non-covalent contacts. Most of these properties are based on shortest paths observable in the graph. The betweenness (*BN*) of a node is defined as the number of shortest paths in the graph passing through that particular node. The term is normalized by the number of node pairs $0.5 \cdot n \cdot (n-1)$ in the protein graph with n nodes [189, 186]. The closeness (*CL*) of a node is defined the inverse of the average path length to any other node. The clustering coefficient describes the surroundings of individual nodes. All adjacent nodes are collected and the number of edges between these n_k nodes is determined. The clustering coefficient (*CC*) of a node is defined as number of edges between its adjacent nodes, divided by the maximum number of edges which can theoretically connect these nodes $0.5 \cdot n_k \cdot (n-1)$. The distinct neighborhood count (*NC*) captures how many sequentially distant (tertiary) protein regions are connected by a residue (see Chapter 5).

6.1.3. DESCRIPTION OF THE GENERALIZED MATRIX LEARNING VECTOR QUANTIZATION CLASSIFIER

The generalized learning vector quantization (GLVQ) is a powerful distance- and prototype-based classification method for class-labeled data[231]. The idea is adapted from unsupervised vector quantization methods such as k-Means or self-organizing maps (SOMs) and an extension of the heuristic learning vector quantization (LVQ) [234]. At least one prototype is initialized for each class and a function, which approximates the classification accuracy (Figure 6.1), is maximized during learning. The optimization is commonly done by stochastic gradient ascent (SGA) and allows for an intuitive adaption of the prototypes. In each iteration, for one training data point $\mathbf{v}$ two prototypes are taken into account: the nearest prototype with the same label as the data point and the nearest prototype with a different label, noted as $\mathbf{w}^+(\mathbf{v})$ and $\mathbf{w}^-(\mathbf{v})$. The correct prototype $\mathbf{w}^+(\mathbf{v})$ is attracted while $\mathbf{w}^-(\mathbf{v})$ is repulsed. The strength of attraction and repulsion is obtained by the gradients of the cost function and the according learning rates. The trained model is a nearest neighbor classifier, i.e. an incoming data point is assigned to the same class as the nearest prototype. In general, the GLVQ is a sparse model with interpretative prototypes. The complexity of the model can be chosen by the user by selecting the number of prototypes per class. If only one prototype per class and the Euclidean distance is applied, the GLVQ is a linear classifier. A more detailed description of the algorithm can be found in [235, 236]. Figure 6.2 provides a graphical representation.

A prominent extension of the GLVQ is the Matrix GLVQ [232]. Beside the prototypes, a mapping of the data points is learned for better separation of the classes (Figure 6.3). This linear mapping, denoted by $\Omega \in \mathbb{R}^{M \times D}$, is powerful and provides additional information about the classification problem. Thereby, D is the number of features. The parameter M

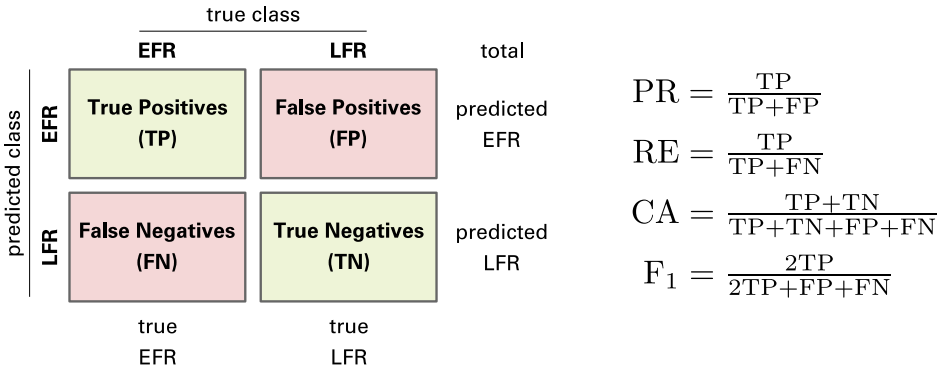

Figure 6.1.: Confusion Matrix and Derived Evaluation Scores

Exemplified Confusion Matrix with the formulas of precision (PR), recall (RE), accuracy (CA), and F_1-measure.

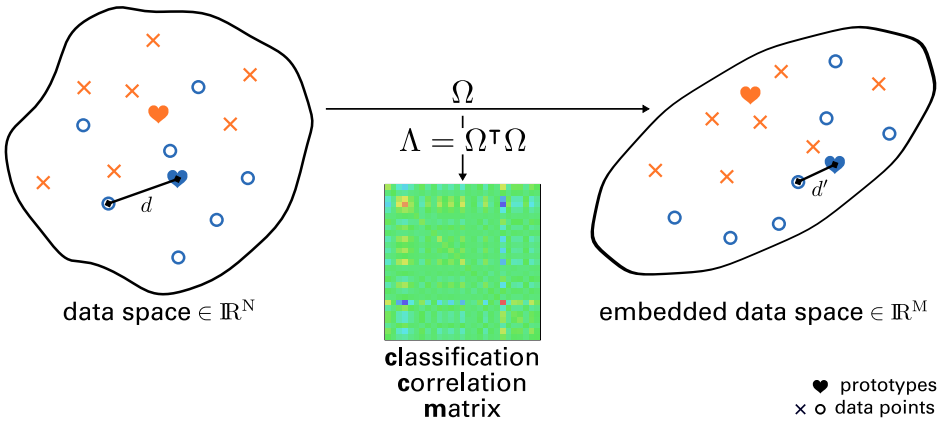

Figure 6.2.: Principle of Generalized Matrix Learning Vector Quantization

Graphical depiction of learning with GMLVQ [228, 229]. One or multiple prototypes represent classes: each data point in the data space of dimension N belongs to the class of the prototype with the closest distance d. Prototypes are updated during learning as in LVQ [237]. Additionally, the matrix Ω maps the data space to an embedded data space of dimension M, where mapped distances d' are optimized. The matrix $\Lambda = \Omega^T\Omega$ (classification correlation matrix) represents the impact of each feature on the classification performance.

can be chosen by the user and indicates the mapping dimension. If the mapping dimension is equal to D, the matrix is quadratic, but M can also be set to values smaller than D, e.g. down to $M = 2$. In the latter case the GMLVQ can be used for visualization of the dataset by mapping the dataset into the two-dimensional space [238]. Moreover, the matrix $\Lambda = \Omega^T\Omega$ is termed classification correlation matrix (CCM) [235]. In contrast to the correlation matrix of the features, the CCM reflects the correlations between them under the aspect of class discrimination (Figure 6.5B). Positive or negative values of high magnitude between two features indicate a high positive or negative correlation of the features for the discrimination of classes (see Figure 6.5A).

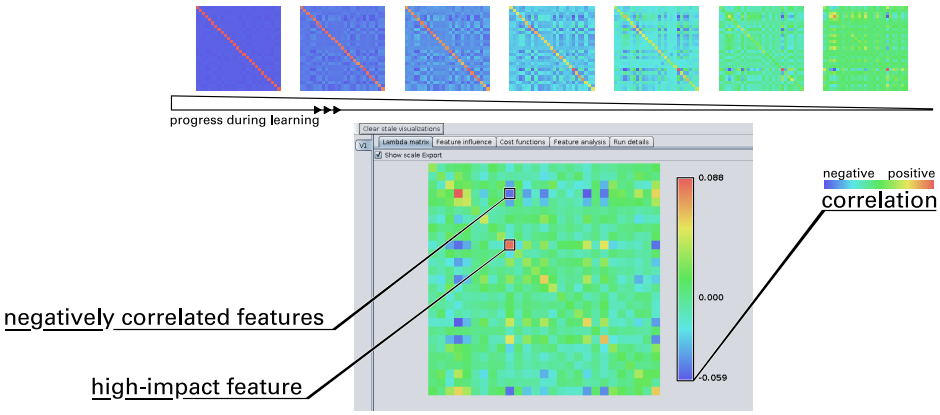

Figure 6.3.: Scheme of the Process of Learning

The graphical user interface of the GMLVQ Weka implementation. The matrix panel shows the CCM and displays live updates during the learning process. A color bar represents the scale of the matrix elements with a coloring scheme similar to a heat map.

6.2. RESULTS & DISCUSSION

6.2.1. CLASSIFICATION OF EARLY FOLDING RESIDUES REVEALS IMPORTANCE OF HYDROPHOBIC INTERACTIONS

The given dataset has a very unbalanced class distribution: 482 data points of class EFR and 2,784 of class LFR. The classification accuracy is inconclusive in such cases because it only takes correctly classified data points into account, e.g. relatively a good classification accuracy would be achieved when all data points were to be classified as LFR. Therefore, we determine further prominent evaluations measures based on the confusion matrix such as precision, recall, F_1-measure, and area under the receiver-operating characteristic (auROC) [239, 240]. The precision considers data points predicted as the positive class (here EFR) and recall on all data points, which are true positives. In the example, the number of EFR is drastically smaller than that of LFR, so in general the precision is worse than recall. The F_1-measure, which is the harmonic mean of precision and recall, is sensitive if one of these values is getting small. Other β values can be specified for the F_β measure. The receiver-operating characteristic (ROC) is a graphical plot illustrating the trade-off between true positives and false positives for a parametrized model. According to the Weka documentation, the ROC is obtained by varying the threshold on the class probability estimates.

We applied 10-fold cross validation on different classifiers to compare the results of GMLVQ to other state-of-the-art methods (see Table 6.2). Furthermore, we investigated different parameter settings of the GMLVQ. On the one side, the model size of the GMLVQ is a parameter chosen by the user. Here, we chose one prototype per class resulting in a linear classifier and five prototypes per class, which is more complex. Moreover, the GMLVQ has the feature to optimize other measures based on the confusion matrix like the F_β-measure or a linear combination of precision and recall. These can take the unbalanced class distribution into account. The comparison of different classification models is challenging. It is hard to decide objectively which classifier performs best. The support vector machine (SVM) ends up with the best accuracy, yet the recall is low. On the other side, the GMLVQ optimizing the weighted accuracy has the best recall and F_1-value and optimizing the F_β-measure ends up with the best value regarding the auROC. Furthermore, we can notice that very complex models do not automatically perform better. The naive Bayes (NB), a simple, fast and, linear classifier performs comparable to the other much more complex models like random forest (RF) or SVM, which utilizes 1,193 support vectors. 36% of the data points are necessary to

describe the hyperplane which indicates a complex model. The GMLVQ with five prototypes per class perform better in training than GMLVQ with one prototype, yet, in test the sparse model is more suitable. Different versions of the cost functions evaluating approximated values of classification accuracy, weighted classification accuracy, F_1-measure, or weighted precision-recall were applied. The results with the according parameter selection are listed in Table 6.2 and Table 6.3.

Table 6.2.: Results of the Learning Process

CM	CA	PR	RE	F_1	auROC
Naive Bayes					
187 | 195 195 | 2,190	72.8	23.9	38.8	29.6	70.9
Random Forest					
192 | 290 293 | 2,491	82.1	39.6	39.8	39.7	64.7
Support Vector Machine					
134 | 348 78 | 2,706	**87.0**	**63.2**	27.8	38.6	62.5
GMLVQ with 1 prototype per class					
run 1					
320 | 162 830 | 1,954	69.6	27.8	66.4	39.2	67.7
run 2					
351 | 162 890 | 1,954	68.7	28.3	**72.8**	**40.7**	73.7
run 3					
348 | 134 891 | 1,893	68.6	28.1	72.2	40.4	**76.6**
GMLVQ with 5 prototype per class					
run 4					
187 | 295 443 | 2,341	77.4	29.7	38.8	33.6	69.4
run 5					
288 | 194 819 | 1,965	69.0	26.0	59.8	36.2	70.5
run 6					
274 | 208 763 | 2,021	70.3	26.4	56.8	36.1	70.3

The test results in % right of the confusion matrix and algorithmic parameters used for the classification of the data determined with Weka. The best values for the single evaluation measured are marked bold. If not stated otherwise, default setup was used. SVM with RBF-kernel ($\sigma = 5$) which results in 1,193 number of support vectors. Weights for weighted accuracy: 0.75 and 0.25. F_β-measure with $\beta = 1$.

To sum up, GMLVQ provided better results in recall even if the model is chosen very sparse. Distinguishing EFR and LFR is challenging and a clear separation was not achieved using the described features. GMLVQ was trained on the dataset in order to retrieve the most discriminative features of EFR and to showcase the capabilities and handling of the visualization.

Table 6.3.: Run Parameters

parameter	run 1	run 2	run 3	run 4	run 5	run 6
cost function to optimize	CA	WCA	F_1	CA	WCA	F_1
number of epochs	150	150	150	250	250	250
number of prototypes	1	1	1	5	5	5
data point ratio per round	0.75	0.75	0.75	0.75	0.75	0.75
sigmoid sigma interval	[1.0,5.0]	[1.0,15.0]	[1.0,50.0]	[1.0,5.0]	[1.0,15.0]	[1.0,50.0]
prototype learning rate	1.0	1.0	1.0	1.0	1.0	1.0
matrix learning rate	1.0	1.0	1.0	1.0	1.0	1.0
omega dimension	27	27	27	27	27	27
cost function beta	-	-	1	-	-	1
cost function weights	-	[0.75,0.25]	-	-	[0.75,0.25]	-
parallel execution	true	true	true	true	true	true

This table presents the parameters used to obtain the results of Table 6.2. Classification accuracy (CA), weighted classification accuracy (WCA) with weights 0.75 and 0.25, as well as F_β-measure with $\beta = 1$ (F_1).

6.2.2. VISUALIZATION OF LEARNING PROCESS AND INTERPRETATION OF CLASSIFICATION RESULTS

The GMLVQ plug-in tracks and summarizes each run by various visualization panels (Figure 6.4): the CCM panel (Figure 6.4A), the cost function panel (Figure 6.4B), the feature influence panel (Figure 6.4C), the feature analysis panel which depicts the prototype placement (Figure 6.4D), and the run details panel which reports the parameters of the corresponding run (Figure 6.4E). A detailed description on the example for the EFR dataset is given in order to demonstrate how results of GMLVQ can be interpreted by integrating information of these visualization panels.

For the presented dataset, the CCM (Figure 6.5A) is primarily homogeneous which is indicated by values close to zero. The major contributing features are the *LF*, *PlipBN*, and especially *PlipHpCL* as these features exhibit the highest scores on the main diagonal of the CCM. The positive correlation of *LF* and *PlipBN* contributes to the classification performance as indicated by positive values described by the corresponding element. Also, the negative correlation of *PlipHpCL* to both features increases classification performance. The *PlipHpCL* is negatively correlated to various other features such as *SecSize*, *PlipLR*, *PlipHbLR*, and *PlipHbCL*. To a lesser degree, *e* and *PlipNC* are associated positively. It has to be

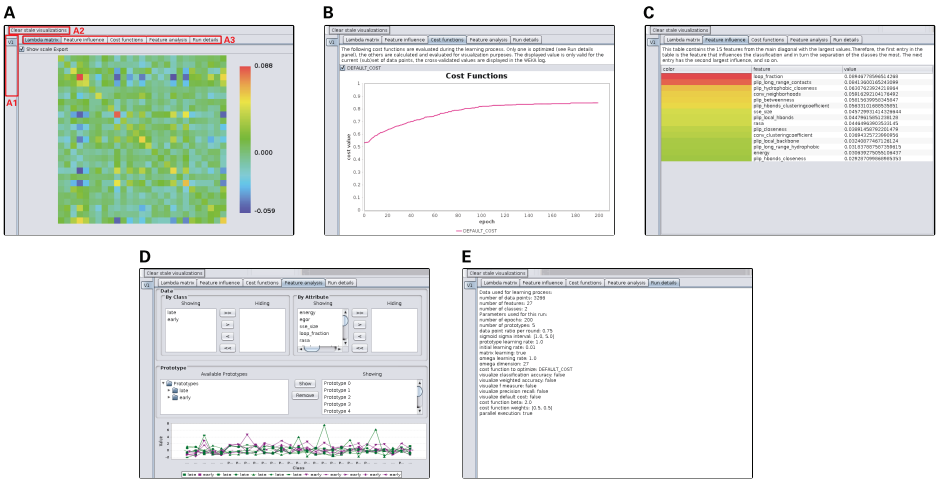

Figure 6.4.: User Interface of the Weka Plug-In

(**A**) The visualization of the CCM. The color scale indicates positive or negative correlations. (**A1**) The visualizations of each separate run will appear in this area. By clicking on the respective tab, one can easily switch between individual runs, e.g. cross validation runs. (**A2**) This button clears all visualizations except the latest. (**A3**) The tabs allow the user to switch between different visualizations of the current run. (**B**) The chart visualizes the earning functions over the course of learning. Additional earning functions can be visualized here, alongside with the cost function which is optimized. (**C**) The feature influence of single features of the current run. The top-ranked features has the highest contribution for the classification performance. (**D**) The feature analysis panel allows the detailed investigation of features and prototypes. (**E**) This panel shows the parameters which were used for the current run.

pointed out that the CCM differs substantially from the correlation matrix (see Figure 6.5B). In the correlation matrix, strong positive correlations are present in the fourth group of features (local contact counts) and negative correlations in the fifth group (tertiary contact counts). Relevant associations between features pointed out by GMLVQ are not obvious from the correlation matrix. The five most important features for discrimination are listed in Table 6.4 which was derived from the feature influence panel (Figure 6.4C). The prototype placement depicted in the feature analysis panel (Figure 6.4D) describes which values individual features adapt for optimal classification performance. This information is not evident from the CCM but necessary for the interpretation of the learned model. Selecting only these five features and learning a model on this dimensionality-reduced dataset, shows a performance similar to the full model. GMLVQ with weighted accuracy and one prototype per class is given in Table 6.5. Recall and F_1 value are even better compared to using all features. Thus, GMLVQ can also be used for feature extraction.

Table 6.4.: Summary of the Top Five Features

feature	influence score
PlipHpCL	0.159
LF	0.127
PlipBN	0.063
SecSize	0.059
e	0.042

These are the most important features for the classification of EFR. The influence score is in arbitrary units, higher values refer to features which increase classification performance.

The homogeneity observed in the CCM is the result of the similarity of several features.

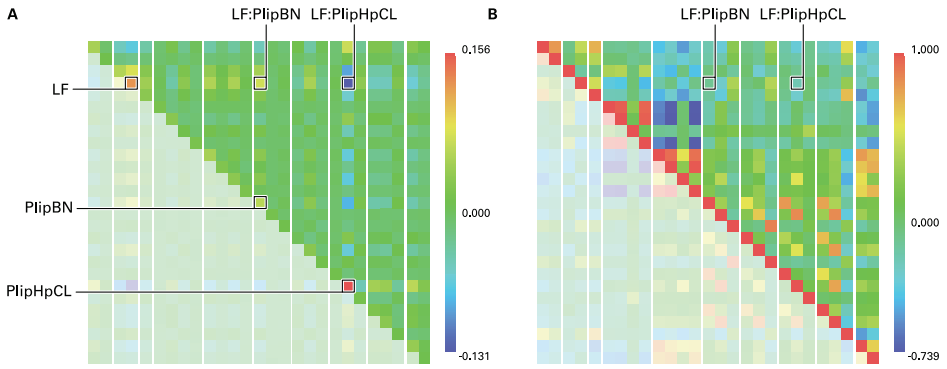

Figure 6.5.: Correlation Classification Matrix for Early Folding Residues

(**A**) The CCM depicts the positive impact of individual features for the classification performance on its main diagonal. Especially, *PlipHpCL*, *LF*, and *PlipBN* are features which discriminate EFR and LFR. The influence of ordered secondary structure elements was shown before [13]. Both betweenness and closeness tend to be increased for EFR which indicates their importance for the assembly of secondary structure elements by tertiary hydrophobic interactions. Other entries of the matrix describe pairs of features which are positively (red) or negatively (blue) correlated and increase classification performance further. (**B**) The standard correlation matrix of all features of the whole dataset. Again, positive and negative correlations are depicted in red and blue respectively. Interestingly, the features pointed out by GMLVQ do not stand out. Vice versa, strong correlations between features do not imply a favorable influence on the classification performance.

Table 6.5.: Performance Using the Five Most Important Features

CM		CA	PR	RE	F_1	auROC
376	106	67.7	28.4	78.0	41.6	69.0
950	1,834					

Classification accuracy (CA), weighted classification accuracy (WCA) with weights 0.75 and 0.25, as well as F_β-measure with $\beta = 1$ (F_1).

At a trivial level, topological descriptors computed on differing graph definitions are likely to result in redundant information. In that case, it is coincidental which feature will be highlighted even though all other correlated features capture similar information. Even if such features are strongly correlated, the CCM will only capture these characteristics if the correlation also contributes to the classification performance.

The *PlipBN* feature is the betweenness [189, 186] derived from all contacts such as hydrogen bonds or hydrophobic interactions [37] in a protein structure. For this residue graph, residues with many of the shortest paths passing through them exhibit high betweenness scores. This feature is highly discriminative between EFR and LFR as captured in the CCM. The prototypes which represent the EFR class display above average *PlipBN* values, indicating that EFR are better connected in the residue graph than their LFR counterparts. In fact, EFR exhibit a higher degree and are crucial connectors, so-called hubs. Residues with high betweenness values have been shown to be crucial for the formation of stable, local structure (foldons) and often constitute the folding nucleus of proteins [186, 94, 208] especially in the defined-pathway model [18, 19].

The *LF* is relatively low for EFR which implies that EFR tend to be surrounded by ordered secondary structure elements. Analogously, this is negatively correlated to the size of the surrounding secondary structure elements and positively correlated to the *Rasa* values as

it has been shown in previous studies [13, 18, 186, 20]. The *LF* feature is furthermore negatively correlated to *e* which indicates that ordered secondary structure elements result in favorable, low energy local structures. These local structures are believed to form autonomously and guide the folding process [18, 19].

The importance of the *PlipHpCL* represents the relevance of hydrophobic interactions in the core of protein structures (Figure 6.6). EFR have an increased propensity to occur in the core of protein structures which is isolated from the polar solvent [87, 13]. However, a buried or exposed state [168] derived from the *Rasa* feature cannot explain the origin and characteristics of EFR. The closeness [190] is defined as the inverse of the average path length of a residue to all other residues. It describes how well-connected individual residues are, which is a similar characteristic as covered by the betweenness [189, 186]. The fact that both *PlipBN* and *PlipHpCL* are the most influential features for the classification demonstrates that they still capture slightly different aspects. The classification performance benefits from a negative correlation of both features. EFR occur primarily in the hydrophobic core of a structure where they participate in an increased number of hydrophobic interactions with surrounding residues. Previously, hydrophobic interactions have been shown to be relevant for the initiation and guidance of the protein folding process itself as well as its *in silico* modeling [241, 242, 27, 243]. Hydrophobic interactions can only be realized by a subset of amino acids and have an increased propensity to form ordered regions [20, 183]. The importance of the *PlipHpCL* feature and the placement of the prototypes implies that EFR are well-embedded in the hydrophobic network of protein structures. EFR may form more hydrophobic interactions which are important for the correct assembly of protein regions separated at sequence level (see Chapter 5).

6.3. CONCLUSION

In summary, the visualized classification of the GMLVQ run pointed out that many features capture redundant information. A subset of the features (*PlipHpCL*, *LF*, and *PlipBN*) is discriminative for both classes. Their importance and their respective correlations are in agreement with previous studies on EFR [11, 13] and, more general, folding nuclei [29, 87, 186, 18, 19].

Machine as well as deep learning are trending in (life) sciences. Yet, a lot of classification problems are difficult to solve. Especially for problems with unbalanced class distributions the choice of the best model is crucial. Beside evaluation measures other properties might be essential to select a suitable classifier. One key aspect is the interpretability of the learning process and the resulting model. GMLVQ is a prototype-based classifier and its applicability was demonstrated on a dataset of key residues in the protein folding process. GMLVQ provides an interpretable classification model and was integrated into the Weka framework to make this classifier and its visualization capabilities accessible to a wide range of scientists.

A dataset of key residues of the protein folding process was investigated. GMLVQ performs comparable to other state-of-the-art methods such as SVM or RF but provides a readily interpretable classification model. From a set of 27 features, GMLVQ identified the fraction of ordered secondary structure elements, the betweenness based on non-covalent contacts, and the closeness using only hydrophobic interactions as the most relevant features for the distinction between EFR and LFR.

The classification performance may be improved by using additional features; however, for sake of simplicity such features were omitted because their computation would require additional algorithms or models. Promising candidates are backbone rigidity values [20], sequence-based predictions of EFR [11], or evolutionary coupling scores [141]. All of them have been previously shown to be discriminative for EFR [11, 13] and may increase the

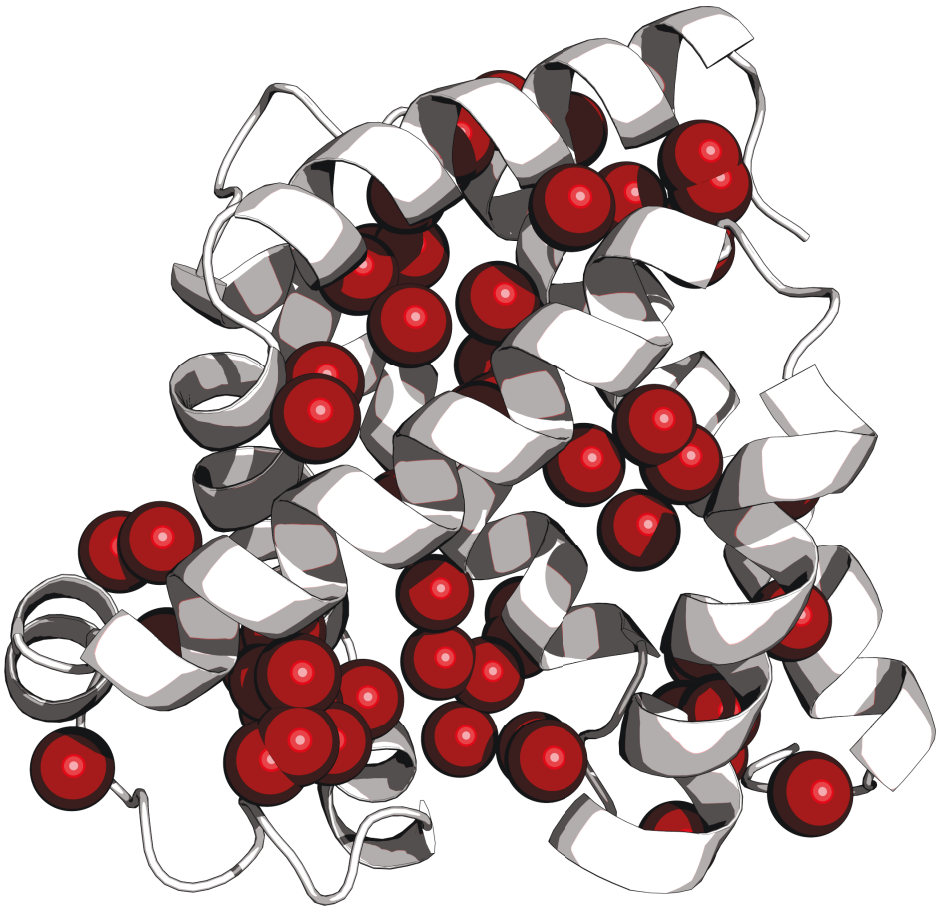

Figure 6.6.: Rendering of the Network of Hydrophobic Interactions

Structure of horse heart myoglobin (PDB:1ymb_A). In this structure, 58 hydrophobic interactions were detected by PLIP [37]. The centroids between interacting residues are depicted as red spheres. This highlights the strong contribution of hydrophobic interactions in the protein core.

classification performance of this exemplary application of the Weka plug-in.

7. THE EVOLUTIONARY HISTORY OF AMINOACYL-TRNA SYNTHETASES

This Chapter is Partially Based on the Publication:

Bittrich, S., Schroeder, M., & Labudde, D. (2018). Characterizing the relation of functional and Early Folding Residues in protein structures using the example of aminoacyl-tRNA synthetases. *PloS one*. 13(10), e206369.

This Chapter is Partially Based on the Publication:

Kaiser, F., Bittrich, S., Salentin, S., Leberecht, C., Haupt, V. J., Krautwurst, S., Schroeder, M., & Labudde, D. (2018). Backbone brackets and arginine tweezers delineate class I and class II aminoacyl tRNA synthetases. *PLoS computational biology*, 14(4), e1006101.

Aminoacyl-tRNA synthetases (aaRS) may be the proteins with the most intriguing evolutionary history and are a prime candidate to analyze as their emergence is well-discussed in literature [244, 245, 246, 247, 248, 2]. aminoacyl-tRNA synthetases (aaRS) enzymes attach amino acids to their cognate tRNA, which is subsequently recognized by its anti-codon and consumed by a ribosome [245]. Thus, aaRS implement the genetic code and give insights into the earliest episodes of life. For each amino acid, a dedicated aaRS implementation exists in each organism [249, 250]. E.g., AspRS attaches aspartic acid to $tRNA^{Asp}$ in two-step reaction which involves the recognition of ATP, amino acid, and tRNA.

$$\text{Asp} + \text{ATP} \xrightarrow{\text{AspRS}} \text{Asp}-\text{AMP} + \text{PP}_i$$
$$\text{Asp}-\text{AMP} + \text{tRNA}^{Asp} \xrightarrow{\text{AspRS}} \text{Asp}-\text{tRNA}^{Asp} + \text{AMP}.$$

Specific aaRS implementations are referred to as type. The 20 types can be divided into two complementary classes (class I and class II, Figure 7.1) with each being responsible for a defined set of amino acids [251, 252, 253]. The amino acids of both classes are distributed evenly regarding their physicochemical properties. However, amino acids handled by class I feature a higher molecular weight. Furthermore, they were shown to occur in the hydrophobic core of protein structures, whereas class II amino acids are more likely exposed to the solvent [254]. Both classes differ significantly at sequence and structure level, feature distinct reaction mechanisms, and occur in diverse oligomerization states – they seem as distinct as possible from each other [248]. It is an open question how both classes emerged and why two strategies to attach amino acids to the proper tRNA molecule emerged. The peculiarities of aaRS suggest the both classes were established at the same point in time and that slight differences of the handled amino acids manifested in the two distinct classes [248]. Over the course of evolution, several domain inserts did occur [247, 248] which render the evolutionary trajectory of aaRS difficult to reconstruct [255]. Some organisms may feature additional aaRS such as PylRS which makes pyrrolysine accessible to protein biosynthesis.

Each aaRS class is defined by a set of sequence motifs which mediate ligand interactions with ATP and realize catalysis [251, 260, 245]. Specific for class I are the HIGH and the KMSKS motifs [251, 245]. Both motifs stabilize the transition state of the chemical reaction. The KMSKS motif occurs as mobile loop in the folded structure [245]. Class II aaRS feature a less conserved [255] set of sequence motifs which are more flexible regarding their relative arrangement [251]. Motif "1" is the youngest motif [2] and realizes dimerization of class II aaRS [245, 261]. Motif "2" and "3" implement the reaction mechanism and encompass two conserved arginine residues [262, 263, 251]. Sequence identity within each class is below 10% [264]. Class I features a structural rearrangement upon ligand binding [265, 244, 266] which stores energy during the reaction in a constrained conformation of the KMSKS motif [267]. No such mechanism is described for class II, but it is assumed that the observed dimerization [245, 261] fulfills a similar role in class II [2]. The catalytic core domain of class I aaRS resembles a Rossmann fold [268, 269], whereas structure of the class II is unique [270, 271, 272]. Also functional differences have been described: class I aaRS attach the amino acid to the 2'OH-group of the tRNA, whereas class II proteins use the 3'OH-group [273]. In summary, aaRS classes share no similarities [245, 272, 248] beside their actual function [244, 247, 248].

In a recent large-scale structural study by Kaiser et al. [2], two ligand binding motifs – the Backbone Brackets and the Arginine Tweezers – were identified as characteristic for each aaRS class. Furthermore, this publication proposes a structure-guided MSA for each class which was shown to be a suitable approach for these highly diverse sets of proteins [2]. Gene fusion, duplication, and recombination events as well as horizontal gene transfer [274, 272] over a period of 4 billion years resulted in sequences difficult to align conventionally.

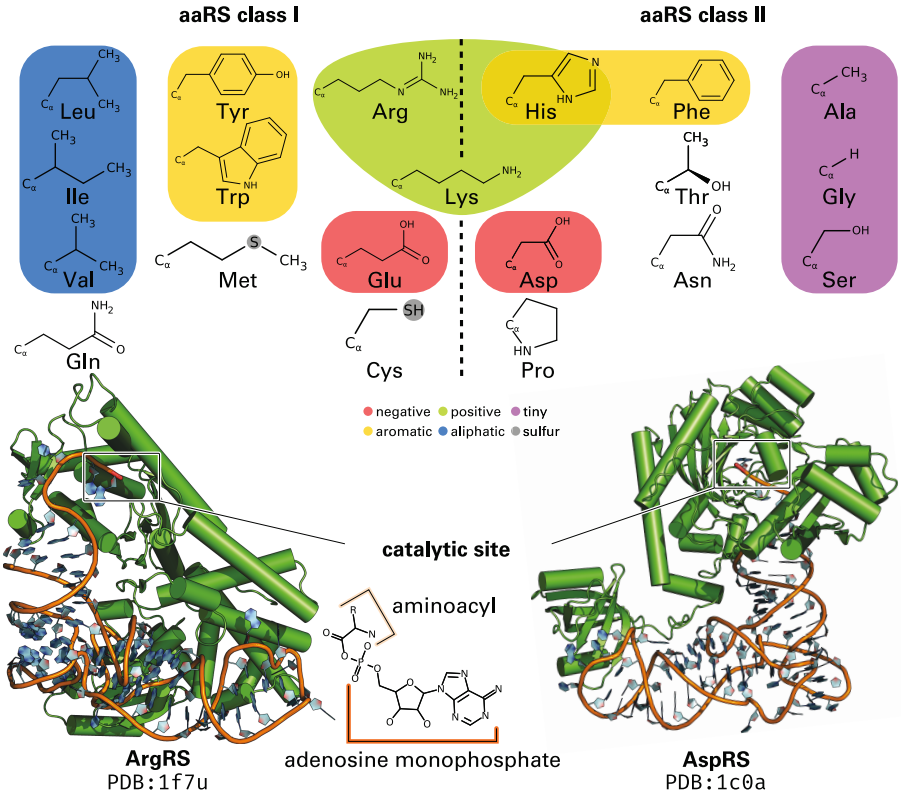

Figure 7.1.: The Two Aminoacyl-tRNA Synthetase Classes and Amino Acids They Handle

The 20 amino acids are distributed evenly between both aaRS classes. Certain physicochemical properties [256] such as aromatic or charged amino acids occur in both classes, so that no clear distinction can be made. Class II handles significantly smaller amino acids [254] and exhibits smaller binding sites [257, 258]. Lysine is mostly processed by class II, however archaic organisms feature a class I aaRS for lysine [259]. Prior to tRNA ligation, the amino acid ligand is converted to its activated form: aminoacyl adenylate. This implies that some residues in aaRS recognize the corresponding amino acid (upper bracket at the aminoacyl moiety) while others recognize ATP (lower bracket).

The Rodin-Ohno Hypothesis Primordial implementations of both aaRS classes called protozymes [247, 248] were linked to the Backbone Brackets and Arginine Tweezers motifs. The Rodin-Ohno hypothesis [244] proposes that aaRS enzymes were once complementarily encoded by the same gene (Figure 7.2). The hypothesis is substantiated by deconstruction of contemporary aaRS sequences. For each class, sequences were aligned and domain inserts were removed which reduced each class to a peptide of 46 residues. These peptides were furthermore manipulated to pair them complementarily, this recreated the hypothesized primordial gene organization. Interestingly, the resulting protozymes have been demonstrated to be catalytically active and enhance reaction rates by orders of magnitudes. Both protozymes resemble molten globules which lack a defined tertiary structure and likely rearrange structurally upon ligand binding. The complementary coding imposes strong constraints on the evolvability of protein sequences [247, 248] and may explain the mirror-like characteristics [275] of both aaRS classes [245, 272, 248]. A similar gene organization was postulated for other protein families as well [276, 277] and may be common to ancient proteins which emerged during a time when size of the genome was strongly limited.

The Rodin-Ohno hypothesis provides an elegant explanation for the emergence and peculiarities of contemporary aaRS classes [244, 247, 248, 2]. It is hypothesized that all aaRS genes originate from this primordial gene encompassing both protozymes. They diverged

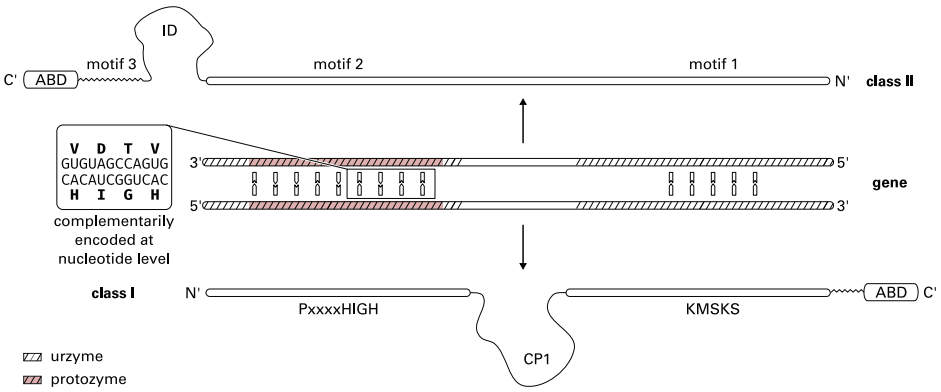

Figure 7.2.: The Rodin-Ohno Hypothesis: Both Aminoacyl-tRNA Synthetase Classes Were Once Encoded on Opposite Strands of the Same Gene

The protozymes of aaRS superfamilies originated from the complementarily encoded region encompassing the HIGH and motif "2" (shaded in red). Contemporary aaRS proteins feature several domain inserts such as the insertion domain (ID), connecting peptide (CP1), and the anticodon binding domain (ABD). Figure adapted from [278, 247].

and improved in specificity, but their catalytic core has been conserved [2].

All of the contemporary aaRS types are connected by the requirement to bind ATP. For the class I protozymes [247], this basal unifying characteristic was found to involve hydrogen bonds. Regardless, aaRS feature an intriguing evolutionary history. Also, the relation of structurally and functionally relevant residues has been studied for class I [279]. Both considerations lead to aaRS being used as an example of two evolutionary diverse protein superfamilies.

7.1. MATERIALS & METHODS

7.1.1. DATASET CREATION

The corresponding dataset was defined as described by Kaiser et al. [2]. Protein structures of aaRS were retrieved from the PDB. Pfam [280] and number of enzyme in Enzyme Commission's system (EC) identifiers were utilized to define aaRS function. Putative chains were excluded. Detailed selection criteria are described in S1 Appendix of the corresponding publication [2]. For each catalytic chain it was determined which ligands are present. Explicitly, amino acid ligands, ATP, and aminoacyl-AMP (the intermediate after the first reaction step) were considered summarized in Figure 7.3 [2].

Some sequences are highly similar to other entries and may introduce some bias into analysis. Therefore, a single-linkage sequence clustering was employed. Sequences were considered similar when their sequence identity according to a Needleman-Wunsch alignment [281] did exceed 95%. If a resulting sequence cluster contained more than one sequence, a representative for the cluster was determined. The selection scheme is described in S2 Appendix of the corresponding publication [2].

As stated, aaRS sequences may be diverse and not comparable directly. A structure-guided MSA by T-Coffee expresso [282] was performed for all 81 representative sequence of class I and for all 75 of class II. The resulting alignments were used to class-specifically renumber all structures in the dataset using a custom script (available as "MSA PDB Renumber" at `github.com/vjhaupt`. All amino acids of the protein chains were renumbered. Atom numbers, chain identifiers, and residue numbers of ligands were unmodified [2].

Subsequently, ligand interactions were annotated by PLIP [37]. All renumbered sequence

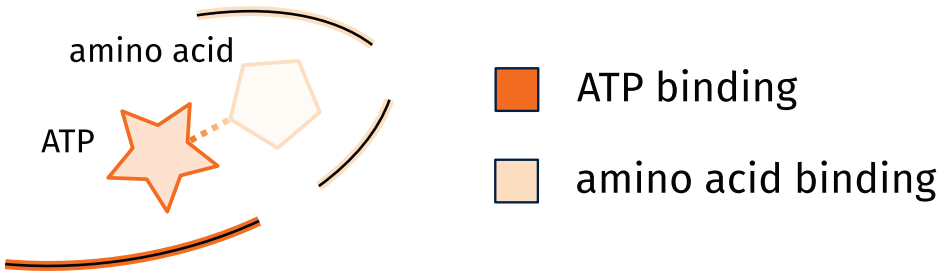

Figure 7.3.: ATP and Amino Acid Binding Site

aaRS enyzmes recognize ATP (dark orange) and amino acid (light orange) ligands. After the first reaction step, they are present as covalently bound aminoacyl-AMP. Interacting residues of the proteins were distinguished between ATP (dark orange line) and amino acid binding (light orange lines) regions. Figure adapted from the publication of Kaiser et al. [2].

positions interacting with a relevant ligand were considered functional. This allowed to identify potential ligand sites in aaRS for structures with no ligand present [2].

7.1.2. PREDICTION OF EARLY FOLDING RESIDUES

From the renumbered protein chains, the corresponding sequence was extracted and used as input for the EFoldMine algorithm [11] which predicts the probability of residues being EFR. This was necessary because no experimentally derived folding characteristics are available for aaRS proteins. Predicted scores exceeding 0.163 where considered EFR; this value has been shown to optimally separate EFR and LFR [11]. Protozyme regions were extracted from PDB:1euy_A and PDB:1c0a_A to represent aaRS class I and II. This selection was for visualization purposes only and focused on structures with aminoacyl-AMP ligands. Selected residue numbers of the protozymes are 255–336 and 648–718. The sequence conservation in aaRS sequences was computed by Jalview [256, 283] using only sequences which were used as input of the MSA. Positions composed of sets of amino acids with similar characteristics result in high values. The referenced MSA was also used to pair both protozyme regions complementarily to create the schematic representation in Figure 7.5. Furthermore, the observed intersection between EFR and functional residues was expressed as probability and compared to the expected probability of a residue to share both the EFR and functional label based on their respective probabilities to occur individually.

7.2. RESULTS & DISCUSSION

A dataset of aaRS structures was published by Kaiser et al. [2]. It encompasses 972 chains containing 448 catalytic domains for class I and 524 for class II. At least one ligand-bound structure exists for each aaRS type. Furthermore, the dataset provides information on ligand interactions, which allow the search for common ligand interaction properties.

7.2.1. BACKBONE BRACKETS AND ARGININE TWEEZERS

Non-covalent protein-ligand interactions were analyzed to investigate the one unifying aspect of aaRS classes: the binding of ATP. Two structural motifs were described by Kaiser et al. [2]. Class I employs backbone hydrogen bonds, whereas two conserved arginine residues furnish salt bridge and π-cation interactions in class II.

The positions in class I exhibit remarkably conserved backbone interactions observable in 441 of 448 (98%) class I structures at renumbered positions 274 and 1361. The nitrogen or

oxygen of the peptide bond interacts with the adenosine phosphate part of the ligand (Figure 7.4A). Both residues enclose the ligand in a bracket-like arrangement (Figure 7.4B) and were thus named Backbone Brackets. Position 274 of the motif shows no conservation at sequence level, whereas position 1361 features preferably hydrophobic amino acids such as leucine, valine, or isoleucine (Figure 7.4C). That both positions are not conserved at sequence level is substantiated by the fact that the side chain of the corresponding residues does not form ligand contacts. The Backbone Brackets occur in unordered secondary structure elements [2].

For class II, conserved interactions are evident at renumbered positions 698 and 1786. Both residues were present in 482 of 524 (92%) of all structures. The Arginine Tweezers motif grasps the adenosine phosphate part of the ligand by salt bridges furnished by the side chain of the respective amino acid (Figure 7.4D), this arrangement resembles a pair of tweezers (Figure 7.4E). Both positions are highly conserved at sequence level (Figure 7.4F). Close to the N-terminal residue of the Arginine Tweezers, a conserved glumatic acid is present at renumbered position 700. At this position a hydrogen bond to the adenine group of the ligand can be observed for SerRS, HisRS, ThrRS, LysRS, ProRS, and AspRS. The N-terminal residue of the Arginine Tweezers predominantly occurs in unordered secondary structure elements, whereas the three positions toward the C-terminus tend to exhibit a β-strand conformation. The C-terminal residue of the Arginine Tweezers at position 1786 occurs in an α-helix. In contrast to the not conserved residues of the Backbone Brackets, both arginine residues are crucial for binding the adenosine phosphate part in aaRS [2]. Mutations to these positions result in a loss-of-function [284, 285].

For both aaRS classes, a stunning balance of evolutionary diversification [286] and equality in function can be observed. This is substantiated by a strikingly different realization of ligand recognition in terms of adjacent sequence positions (Figure 7.4C and Figure 7.4F). The Backbone Brackets motif was described by Kaiser et al. [2] for the first time. Both structural motifs add a new level of profound differences between both aaRS classes not identified before [2].

The relevance of backbone interactions are often underestimated in structural studies. However, backbone hydrogen bonds account for at least one quarter of ligand hydrogen bonding [288]. As long as the backbone orientation is correct, properties of the side chain such as steric effects play a subordinate role. Backbone hydrogen bonds are crucial for NAD binding in a CysG protein from *Salmonella enterica* (PDB:1pjs) as demonstrated by PLIP [37]. Together with the low sequence conservation of the Backbone Brackets motif, this emphasizes that the Backbone Brackets are a relevant example of conservation at function level [2]. Sequence-based analysis cannot spot all relevant aspects of proteins and neither can mere structure analysis. Rather different levels of information have to be integrated to get a more complete picture needed to understand proteins. In accordance with the functionalist principle, function is conserved over structure or sequence [4]. An advantage of the Backbone Brackets motif is that it is resilient to mutations as there are virtually no limitations on the amino acids which can realize such backbone hydrogen bonds [2]. Complementary coding [244, 247, 248] imposes strong constraints on the sequence of both protozymes. Potentially the Backbone Brackets motifs emerged because it eases the evolutionary pressure on certain positions of the class I protozyme. Any amino acid can furnish the observed backbone hydrogen bonds to the ATP ligand. This drastically increases the evolvability of both protozymes. If the strongest constraints during evolution were imposed on protein structure, the evolutionary progress might have been considerably slower [4].

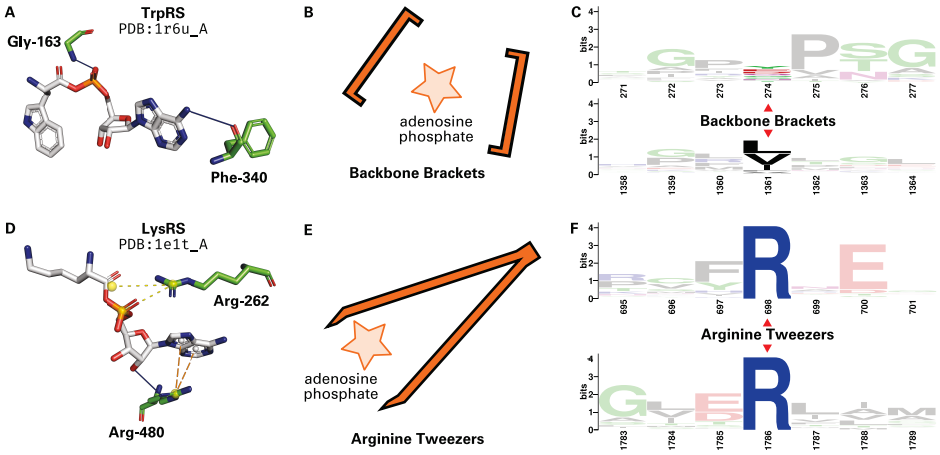

Figure 7.4.: Comparison of Backbone Brackets and Arginine Tweezers

(A) Structural rendering of the Backbone Brackets motif interacting with Tryptophanyl-5'AMP ligand (TrpRS, PDB:1r6u_A). Ligand interaction is furnished by backbone hydrogen bonds (solid blue lines). Residue numbers are given in accordance to the original structure. (B) The Backbone Brackets motif encircling the ligand in a fashion resembling a bracket. (C) The sequence of Backbone Brackets residues (274 and 1361) and three surrounding sequence positions represented as WebLogo [287]. Renumbered residue positions are given. (D) Structural rendering of the Arginine Tweezers motif interacting with Lysyl-5'AMP ligand (LysRS, PDB:1e1t_A). The ligand is bound via salt bridges (yellow dashed lines) as well as π-cation interactions. Residue numbers are given in accordance to the original structure. (E) The Arginine Tweezers grasp the ligand in a manner comparable to a pair of tweezers. (F) The sequence of Arginine Tweezers residues (698 and 1786) and three surrounding sequence positions represented as WebLogo [287]. The Backbone Brackets show little conservation at sequence level since backbone interactions can be established by all amino acids. In contrast, the Arginine Tweezers rely on two arginine residues to furnish their highly specific interaction types.

7.2.2. THE POSITION OF EARLY FOLDING RESIDUES IS CONSISTENT IN AMINOACYL-TRNA SYNTHETASES

In Chapter 5 a separation of EFR and functional residues can be observed. However, no analysis of EFR in an evolutionary context is feasible due to limitations of this dataset. aaRS may be the protein superfamily with the most intriguing evolutionary history and, thus, are a prime candidate to analyze in the context of the previous findings as their emergence is well-discussed in literature [244, 245, 246, 247, 248]. The Rodin-Ohno hypothesis [244] proposes that aaRS enzymes were once complementarily encoded by the same gene (Figure 7.5). This provides an elegant explanation for the emergence and peculiarities of contemporary aaRS classes [244, 247, 248]. It is hypothesized that all aaRS genes originate from this primordial gene encompassing both protozymes. They diverged and improved in specificity, but their catalytic core has been conserved.

Further analysis focuses on regions of today's aaRS structures which correspond to the protozyme regions in order to assess how EFR predicted by EFoldMine [11] related to functional residues in an evolutionary context. ATP and amino acid recognition sites were considered functional (see Figure 7.3). Furthermore, it was assessed whether the predicted positions of EFR are consistent in these highly diverse superfamilies of enzymes. This analysis is backed by a manually curated dataset which accounts for high diversity of contemporary aaRS implementations.

Fig 7.6 depicts the protozyme [247, 248] of each aaRS class with an aminoacyl-AMP ligand present, which captures the intermediate of the enzymatic reaction.

Analysis is based on 81 non-redundant structures for class I and 75 for class II, respec-

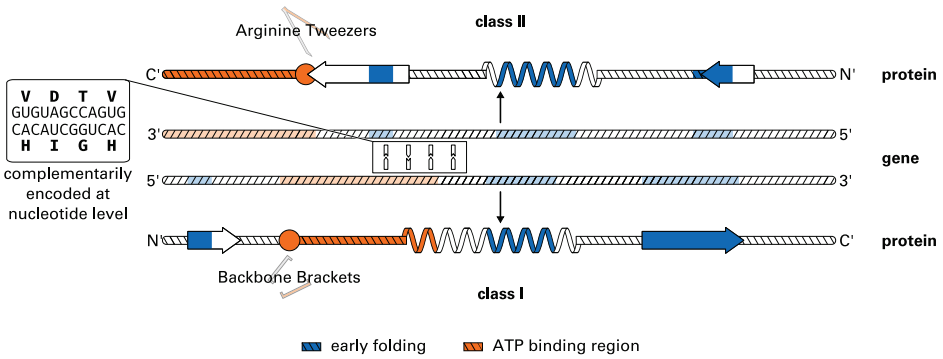

Figure 7.5.: Schematic Representation of Protozyme Regions

The two classes of contemporary aaRS enzymes may originate from opposite strands of the same gene. The corresponding peptides (called protozymes) have been shown to be catalytically active [244, 247]. The order of secondary structure elements in both protozymes resembles a mirror image. Using the EFoldMine classifier [11], EFR (i.e. folding initiation sites) were predicted (depicted in blue). EFR are a distinct set of residues with respect to ATP binding sites (orange) identified in Chapter 7. Backbone Brackets and Arginine Tweezers are class-specific ATP binding motifs identified in the same study. Regardless of aaRS class, EFR occur in the center of secondary structure elements. Their position is preserved within aaRS classes despite sequence conservation being relatively small. The relative arrangement of EFR in class I resembles a prominent structural packing motif [269]. The more general Start2Fold dataset [8] is used to assess whether the separation of EFR and functional residues is a common theme in protein structures. The ATP binding region contains four binding residues each and was simplified to a continuous region for visual simplicity. Figure adapted from [278, 247, 2].

tively. For each analyzed structure the corresponding sequence was used to predict the position of EFR [11]. A consistent numbering of residues within each class was established by a structure-guided MSA [282]. Even within the depicted catalytic core of aaRS structures, sequences feature a high degree of variability and various inserts. Interestingly, residues predicted to be early folding are located at MSA columns which may not be extraordinarily conserved but are present in at least half of the corresponding sequences. EFR positions are mostly conserved among aaRS homologues. ATP binding sites are also consistent for the structures, whereas the position of amino acid binding sites varies. In the visualized protozyme regions (Fig 7.6), positions of EFR are located in ordered secondary structure elements. Functional residues, especially those realizing ATP recognition, are located in spatial proximity to one another. Furthermore, they occur in unordered coil regions and are located close to the ligand. ATP binding sites (dark orange) can be found on the left in proximity of the adenine part, whereas amino acid recognition sites (light orange) can be found on the right close to the amino acid part of the ligand. Sequence conservation scores were considered. For comparison, the highly conserved N-terminal arginine of the Arginine Tweezers motif exhibits a score of 11, scores close to 0 indicate no conservation. The average sequence conservation of the protozyme regions is 1.59 (1.42) for class I (and class II respectively). Positions predicted to be EFR exhibit scores of 2.50 (2.80). That for ATP binding sites is 3.75 (3.75) and for amino acid binding sites 1.85 (2.17). On average the EFoldMine prediction is 0.09 (0.09) for the protozyme regions. Positions considered EFR exhibit high values of 0.21 (0.20). ATP binding sites feature low scores, whereas amino acid binding sites feature slightly increased probabilities of being EFR (summarized in Table 7.1). Because the position of amino acid binding sites is not consistent in the MSA, sequence conservation of these positions is relatively small. In contrast, ATP binding sites are mapped consistently in the MSA for both aaRS classes.

EFR exhibit smaller sequence conservation scores than ATP binding sites which indi-

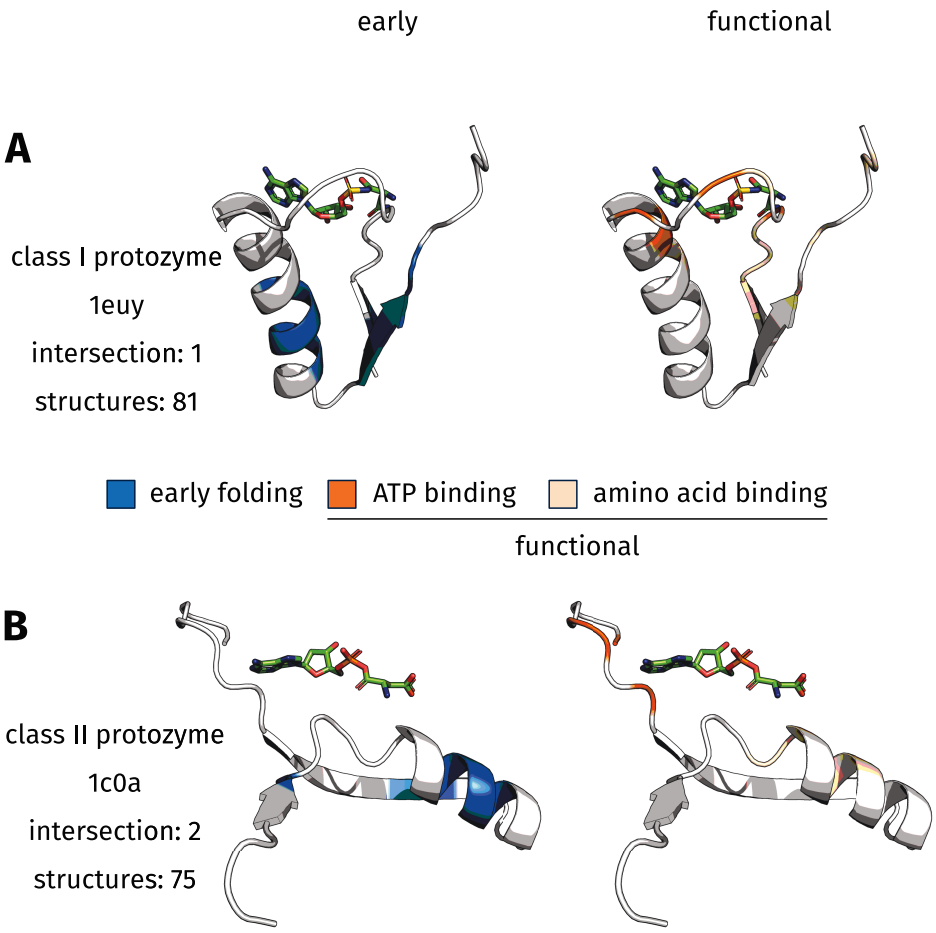

Figure 7.6.: Protozyme Regions of Both Aminoacyl-tRNA Synthetase Classes

The protozyme regions [247, 248] (in cartoon style) and the respective aminoacyl-AMP ligand (in sticks style) are depicted. This captures the state after the first reaction, when ATP and amino acid have been covalently bound. The ATP part is oriented to the left, whereas the amino acid is located on the right. Residues predicted to be early folding [11] are colored blue, whereas functional residues are rendered in orange. ATP interaction sites are depicted in dark orange, residue positions observed to interact with the amino acid in any aaRS structure are rendered in light orange (see Figure 7.3 for a schematic depiction). In the rare cases that residues are both EFR and functional, they bind the amino acid part of the ligand in two specific aaRS implementations. (**A**) The class I protozyme is represented by truncated PDB:1euy_A. The respective EFR are located in the center of the ordered secondary structure elements. In contrast, functional ligand binding sites are located in the upper part of each subfigure. They are primarily located in unordered coil regions. (**B**) The class II protozyme, represented by truncated PDB:1c0a_A, shows similar tendencies.

cates that more sequence variability can be tolerated for folding initiation sites. Protein function depends on particular amino acid side chains [25], whereas protein structure and secondary structure element formation is mainly the consequence of the hydrophobicity of amino acids [42, 43]. ATP binding sites exhibit lower EFR prediction scores compared to the average in the protozyme region which captures their tendency to occur in exposed, unordered coil regions as observed in the previously reported findings.

Table 7.1.: Sequence Conservation and Average EFoldMine Scores for Aminoacyl-tRNA Synthetase Classes

	feature	all	protozyme	early	functional	ATP	aa
class I	conservation	0.17	1.59	2.50	2.29	3.75	1.85
	EFoldMine score	0.08	0.09	0.21	0.08	0.04	0.09
class II	conservation	0.18	1.42	2.80	2.80	3.75	2.17
	EFoldMine score	0.11	0.09	0.20	0.08	0.06	0.09

Sequence conservation [256, 283] and EFoldMine [11] predictions for the aaRS protozyme regions [247, 248] are presented. Encompassed are the average values for all residues, residues in the protozyme region, for positions predicted to be EFR, functional residues, ATP binding residues, and amino acid binding sites.

7.2.3. STRUCTURAL PACKING MOTIF IN CLASS I AMINOACYL-TRNA SYNTHETASES

The LFR position 284 features a remarkably high sequence conservation of 10. This position is part of the HIGH sequence motif which relates to ATP binding and the stabilization of the transition state [245]. In most class I aaRS, the HIGH motif is located at the N-terminal end of an α-helix. This particular arrangement is commonly observed for ATP binding proteins due to the favorable interactions between the negative charge of phosphate moieties and the dipole moment of the helix backbone [289]. Despite the defined secondary structure in this region, the HIGH motif is predicted to consist of LFR. EFR are located close to the C-terminal end of the helix (Fig 7.6A). Such folding initiation sites will lead to an extension of the nascent secondary structure element until certain sequence compositions terminate the process [9, 10]. Within this secondary structure element crucial for function, residues initiating its formation and residues binding the ligand occur at distinct positions. Furthermore, the observed C-terminal aggregation of EFR and the proximity to other EFR in neighboring β-strands substantiates a previously described structural packing motif in the catalytic core of class I aaRS. It is one of the most ancient and most widely distributed structural motif and was identified in a diverse set of proteins which encompasses the catalytic domain of class I, the anti-codon-binding domain in class II, and five other members of the Rossmanoid family [269]. This motif has been associated to a structural rearrangement important for function [290, 291]. The nearby Backbone Brackets motif rearranges upon ligand binding which implies that the structural rearrangement observed is a feature common to all class I aaRS structures (see Chapter 7).

7.2.4. EARLY FOLDING RESIDUES ARE NON-FUNCTIONAL IN AMINOACYL-TRNA SYNTHETASES

In class I (visualized by truncated PDB:1euy_A), position 311 is the only residue which is both EFR and functional (Table 7.2). This position is only functional in TrpRS and TyrRS where it realizes binding of the respective amino acid. Both tryptophan and tyrosine are large, aromatic amino acids and it is hypothesized that they were added to the genetic code recently [246]. This implies that these EFR became functional late during the evolution of aaRS. The clear separation with respect to ATP recognition implies that the unifying aspect of all aaRS is binding of the ATP ligand and catalysis at the respective α-phosphate. At first protozymes where required to bind ATP and later the amino acid binding sites improved in specificity, allowing them to discriminate between amino acids more reliably. Position 274 corresponds to the N-terminal residue of the Backbone Brackets structure motif. Close to this position various amino acid binding sites can be observed in other class I aaRS,

while EFR are further away (Figure 7.6). Despite being functionally relevant, the sequence conservation of position 274 amounts to 3 and is relatively small. This residue has been shown to realize ATP binding by backbone hydrogen bonds which can be virtually realized by all amino acids. Thus, change can be compensated at this position as along as the backbone atoms can still bind the ATP ligand. Furthermore, this position interacts with the α-phosphate position of the ligand to which the aaRS attaches the proper amino acid. Therefore, it is intuitive that many positions involved in amino acid recognition are located at neighbored sequence positions. In class I, 15 of 16 EFR positions in the MSA relate to well-mapped positions (i.e. present in >50% of aligned sequences).

Table 7.2.: Comparison of Folding Characteristics and Functional Relevance for Aminoacyl-tRNA Synthetase Classes

class	early	ATP	aa	ATP int.	aa int.	ATP shift	aa shift
class I	16	4	13	0	1	-0.95%	-1.87%
class II	10	4	8	0	2	-0.82%	1.22%
	26	8	21	0	3	-0.90%	-0.39%

ATP refers to the number of ATP binding sites and aa refers to the number of positions realizing amino acid recognition in any aaRS implementation. The intersection of functional residues involved in ATP and amino acid binding is given. The shift in probability to the expected intersection is stated. A perfect separation of EFR and functional residues in the sense of ATP binding positions can be observed. Also, positions relevant for amino acid specificity are remarkably well-separated from EFR most of the time. The overlap is present in the amino acid recognition sites in two implementations respectively: TrpRS and TyrRS in class I and AspRS and PylRS in class II.

In class II, positions 665 and 666 are both functional and predicted to be EFR (Table 7.2). Again, these positions are not functional in most class II aaRS. Only in AspRS and PylRS they are observed to bind the amino acid part of the ligand. In agreement with the observation for aaRS class I, asparagine and pyrrolysine are relatively large ligands which may require EFR to participate in protein function. 9 of 10 EFR positions are well-mapped in class II. For both classes, functional positions are well-mapped too. For position 698 of class II a sequence conservation score of 11 is observed. This position is the N-terminal residue of the Arginine Tweezers motif which has been demonstrated to depend on the conservation of this amino acid for ATP binding via salt bridges and π-cation interactions. Like in class I, ATP binding positions can be found accumulated together at sequence level without any EFR between them (Figure 7.5). In summary, the position of folding initiation site is preserved in aaRS despite their large evolutionary divergence. Potentially, aaRS even had influence on the organization of the genetic code and may have caused a shift in the interpretation of genetic information. Amino acids handled by class I more often constitute the hydrophobic core of proteins, whereas amino acids handled by class II are more likely to occur at the interface to the polar solvent [254].

7.3. CONCLUSION

The structures of the aaRS superfamilies were analyzed. It is shown that the position of folding initiation sites is preserved over the course of evolution even when the corresponding sequence conservation is small. Folding initiation sites occur in the center of secondary structure elements, independent of aaRS class. Furthermore, the findings related to the protozymes of aaRS substantiate that protein function can be considered the most important aspect of a protein [4] and retaining protein fold may be of subordinate importance [218]. Functional residues (i.e. ATP binding sites consistently shared by all aaRS

types) exhibit a higher sequence conservation than EFR. EFR and functional residues are distinct sets of residues when amino acid binding positions are ignored which are only relevant in a small number of implementations. Even when these amino acid binding positions are considered to be functional in all implementations, the intersection is remarkably small for class I. In both aaRS superfamilies, EFR are located consistently in the same columns of the respective MSA which agrees with the observation that this characteristic depends on the composition of local sequence fragments [11] and is relatively insensitive to inserts.

Regarding the origin of aaRS, the Rodin-Ohno hypothesis states that the peculiar nature of the two aaRS classes is the result of their respective protozymes [247, 248] being encoded on opposite strands of the same gene. Backbone Brackets and Arginine Tweezers were traced back to these protozymes and their more efficient successors, the urzymes. Both structural motifs can be observed as pairs of residues in contemporary structures and it seems that the time of their addition, indicated by their placement in the ancient aaRS, coincides with the evolutionary trace of proto- and urzymes. The designed approach was used to analyze aaRS from the different viewpoints: sequence backed by structure information, and ligand interactions of essential ligand binding patterns. Additionally, the largest manually curated dataset of aaRS structures including ligand information available to date is provided. This can serve as foundation for further research on the essential mechanisms controlling the molecular information machinery, e.g. investigate the effect and disease implications of mutations on crucial binding site residues as well as evolutionary aspects [2].

8. CONCLUSION

This thesis converged on the role of early folding residues in protein structures. Their structural relevance was assessed by a novel algorithm which universally provides an unmatched granularity for the interpretation of contact maps and has implications for protein structure prediction methods.

> **Results For Question I**
> The StructureDistiller algorithm quantifies the structural relevance of individual entries in a contact map.

Knowledge of the Most Relevant Contacts Increases Reconstruction Fidelity A reference implementation for the assessment of the structural relevance of individual contacts is presented. Normally, the role of contacts for structural integrity cannot be assessed because of their context-specificity: rather their influence depends on complex interplay with other residues [14, 15]. Experimental data on the influence of certain contacts is sparse. For cytochrome c is has been observed that a disruption to a hydrogen bond will lead to a loss of structure, wherein the protein will adopt a molten globule conformation. This contact connects two Ω-loops and the information is crucial to correctly associate two otherwise unstructured parts of the protein [178, 179]. The proposed StructureDistiller algorithm correctly identifies contact information between both Ω-loops to be of outstanding importance for structural integrity.

Secondary structure information was provided to the reconstruction algorithm, thus implicitly providing information on local hydrogen bonds. This may explain why hydrophobic interactions are of high structural relevance: they provide information not captured by the restrictions of the covalently bound chain or secondary structure elements. Knowledge of the most relevant contacts increases reconstruction fidelity significantly by 0.4 Å. On a more general note, StructureDistiller allows to interpret contact maps in detail. This may help further leveraging contact maps as stepping stone between protein sequence and structure (Figure 8.1).

StructureDistiller Allows Improved Resilience to False Positive Contacts Contact maps are highly sensitive to false positive predictions: contacts not present in the native structure quickly dilute the reconstruction fidelity which can be achieved using a contact map [16, 17]. The combination of the most relevant contacts as identified by StructureDistiller compensates up to 6% of false positive contacts before the reconstruction quality is worse than that of a random selection of contacts without any false positive contacts. This knowledge demonstrates that the structural relevance of contacts is not uniform. Rather utilizing knowledge of the most relevant contacts has beneficial effects and should be incorporated into contact prediction methods [140, 135].

Highly Stable Residues Are of Outstanding Structural Relevance Highly stable residues resist serious unfolding pressure in the native structure [161]. These residues are assumed to stabilize secondary structure elements (in contrast to early folding residues which promote their formation). Highly stable residues were found to exhibit significantly increased structural relevance scores: when these contacts are deleted from a protein structure, the protein will likely lose its defined structure. This implies that early folding residues are needed for the initiation of the folding process, but not necessarily relevant for ensuring structural integrity. This opens a new avenue for the interpretation of protein structures.

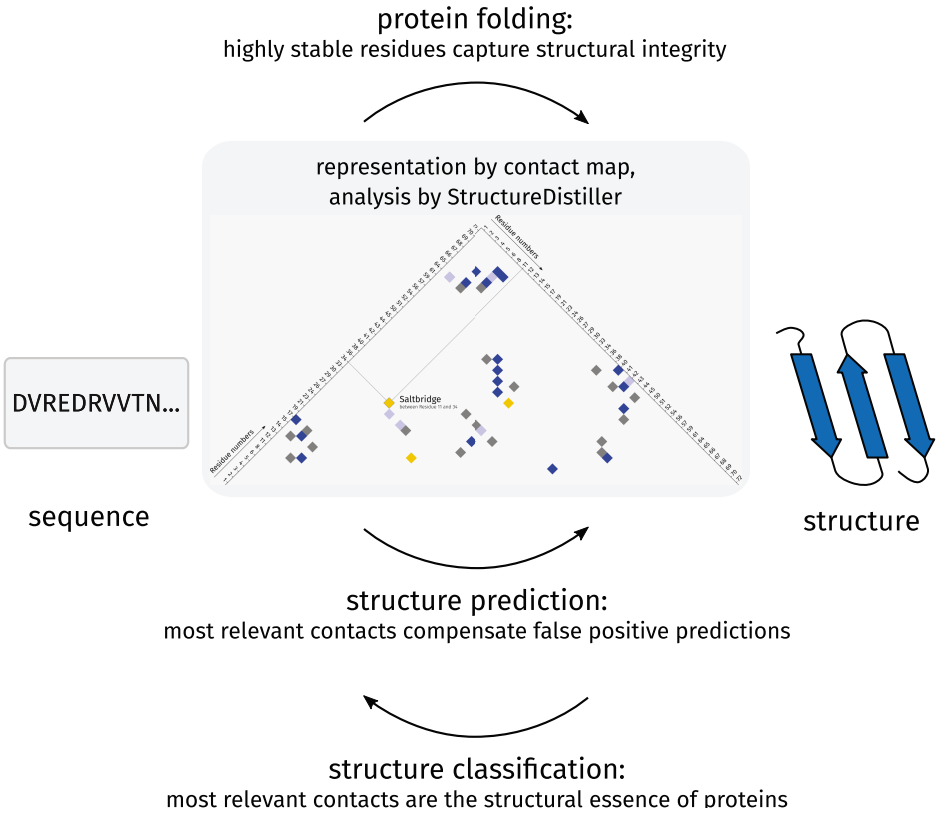

Figure 8.1.: Insights by Structural Relevance Scores

Contact maps are a valuable link between protein sequence and structure with implications for protein folding, structure prediction, and structure classification. The presented StructureDistiller algorithm allows to assess the structural relevance of individual contacts in a contact map and provides a new level of information necessary to interpret them. Knowledge of the most relevant contacts allows to assess the role of early folding residues and highly stable residues, improves structure prediction techniques, and provides a novel way to identify the structural essence of a protein. This last problem was assumed unsolvable a decade ago [14].

Results For Question II

Early folding residues in class I aminoacyl-tRNA synthetases occur in an ancient and widely distributed structural packing motif.

The Positions of Early Folding Residues Are Preserved over the Course of Evolution in Aminoacyl-tRNA Synthetases Early folding residues are not annotated experimentally for diverse protein families; however, they can be predicted from protein sequences with reasonable accuracy [11]. It was demonstrated that even in an evolutionary distant set of proteins (such as aminoacyl-tRNA synthetases) the predicted positions of early folding residues are preserved through evolution. Commonly, the center of secondary structure elements constitutes the folding core. Early folding residues occur at positions for which the amino acid may change over the course of evolution. However, these positions are not prone to insertions or deletions as indicated by the multiple sequence alignment of Kaiser et al. [2]. Thus, early folding residues are linked to structurally conserved regions of proteins, whereas functional residues are commonly embedded in flexible coil regions which tend to contain inserts. This suggests a separation of functionally and structurally relevant residues in protein structures. Protein function can be broken down to a small number of residues with specific characteristics such as reactive hydroxyl groups [25, 221].

In contrast, key residues for folding initiation and structural stability are caused by the hydrophobicity of an amino acid [42, 43, 11]. Several levels of modular characteristics can be observed in proteins [23, 188, 212]: it is intuitive that functionally and structurally relevant residues are distinct entities. This would allow to change protein function without compromising the protein fold. This may increase the evolvability and robustness of proteins after gene duplication events [24]. In particular, this trend may be present in aminoacyl-tRNA synthetases, although their evolutionary history is still too little understood to draw a final conclusion. Furthermore, the conservation of folding initiation sites provides an explanation on how unrelated sequences can adopt a similar fold [5].

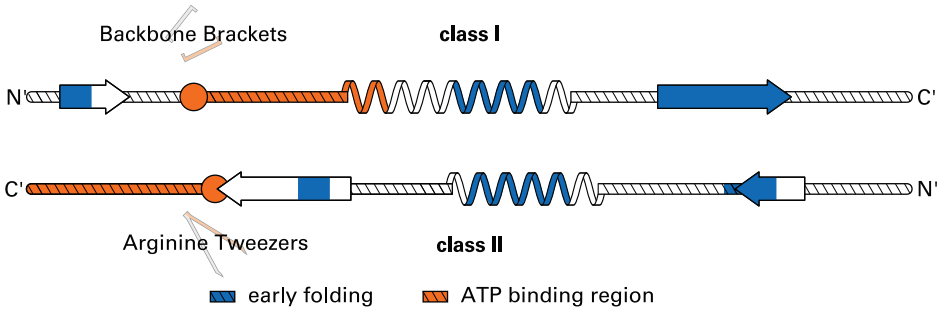

Figure 8.2.: Distribution of Early Folding Residues in Aminoacyl-tRNA Synthetases

Two protozymes have been proposed as primordial implementations of aminoacyl-tRNA synthetases [247]. Therein, the position of early folding residues (blue) is conserved despite 4 billion years of evolution: independent of class, early folding residues occur in the center of secondary structure elements. Functional regions (orange), which bind ATP, occur in unordered secondary structure elements. Early folding residues are in spatial proximity, the arrangement in class I resembles an ancient, widely distributed structural packing motif shared by a diverse set of proteins [269].

The Functionalist Principle in Aminoacyl-tRNA Synthetases Two structural motifs were identified in the highly diverse superfamilies of aminoacyl-tRNA synthetases, which break down their ATP binding function to a pair of residues each. Despite the divergent evolutionary history, contemporary structures use the structurally conserved Backbone Brackets and Arginine Tweezers motifs to bind ATP ligands. 81 and 75 non-redundant structures were analyzed for class I and class II, respectively. The primordial protozymes of both classes [247] imply a structural reductionism as well, because they constitute ancient implementations which feature a drastically decreased number of residues while still allowing the enzymes to retain their biological function. The functionalist principle states that the largest evolutionary pressure is imposed on protein function; function is more conserved than structure, which in turn is more conserved than sequence. In aminoacyl-tRNA synthetases, functional positions exhibit higher sequence conservation scores than folding initiation sites. In class I, the arrangement of early folding residues resembles a structural packing motif that has been identified previously. This motif is assumed to be one of the most ancient and most widely distributed structural motifs. Interestingly, it can not only be observed in the catalytic core of class I, but also in the anticodon binding domain of class II as well as in several proteins of the Rossmanoid family [269].

Hydrogen Bonds Furnish Local Structures – Hydrophobic Interactions Furnish Tertiary Structure The presented findings imply that the role of hydrogen bonds and hydrophobic interactions in the context of protein folding cannot be dissected. Hydrogen bonds form local structures [40, 41] which drastically limit the number of possible conformations of a protein. This may accelerate or even be a strict requirement for protein folding [27]. Contrarily,

hydrophobic interactions furnish tertiary contacts between sequentially separated protein parts. They seem to impose an intrinsic urge on proteins to collapse to a compact formation, though additional information is needed for the correct assembly [21]. Hydrogen bonds can be related to protein function, but the opposite is true for hydrophobic interactions.

Early Folding Residues Are Wired Distinctively by Hydrophobic Interactions Early folding residues provided by the Start2Fold database [8] were analyzed systematically. They exhibit several distinct characteristics as demonstrated by network analysis and the Energy Profiling approach [183]. Interestingly, early folding residues show a significant change (in a bias-corrected population) compared to functional residues in the dataset. Early folding residues prefer ordered secondary structure elements, smaller relative accessible surface area values, and an increased number of hydrophobic interactions. The characteristics of early folding residues also manifest themselves both in sequence and structure as shown by the Energy Profiling approach [183], wherein they constitute more stable local conformations. The application of an interpretable classification model substantiates these findings despite employing a completely different philosophy to determine the most discriminative features. Regarding evolutionary information, functional residues show a significant increase in sequence conservation compared to early folding residues. In consequence, they are more conserved at the sequence level than average residues but exhibit more degrees of freedom than functional residues.

Outlook: On Sequence, Structure, and Function in an Evolutionary Context The relation of protein sequence, structure, and function is difficult to assess, but at the heart of the protein folding problem. Understanding how these aspects are associated – especially in the context of evolution – makes their connection more tangible. This thesis demonstrates that a small number of early folding residues initiate the folding process in proteins; subsequently allowing a small number of functional residues to implement enzymatic reactions, bind ligands, or propagate signals. Nevertheless, the complexity of a protein cannot completely be reduced to specific residue positions because most aspects are context-specific [29]. Larger sequence fragments are needed to manifest the positions of early folding residues and functional site may depend on their surroundings to modulate the binding affinity to a potential ligand or realize structural rearrangements. However, for the first time it was demonstrated how evolutionary forces act distinctively on the function and the structure of proteins. Ultimately, the protein sequence is molded by the necessity for change at these two levels, because it is the only level where change can manifest and be passed down to ancestors. Understanding how evolution affects structurally and functionally relevant regions of proteins is a necessity to tackle e.g. fast-evolving viruses such as the human immunodeficiency or hepatitis C virus [26].

Part III.

CODE AVAILABILITY

Several git repositories were created to address questions during the project. Most underlying source code was made publicly available.

- The eQuant web server can be accessed at:

  `https://biosciences.hs-mittweida.de/equant/`

- Data integration, computation of Energy Profiles, implementation of the StructureDistiller algorithm was realized by a Java library publicly available at:

  `https://github.com/JonStargaryen/jstructure`

- An open-source release and a step-by-step guide for the usage of the GMLVQ plug-in presented in Chapter 6 is available at:

  `https://github.com/JonStargaryen/gmlvq`

- A persistent release of the GMLVQ plug-in is associated to DOI:

  `https://doi.org/10.5281/zenodo.1326268`

- The compiled StructureDistiller algorithm is associated to DOI:

  `https://doi.org/10.5281/zenodo.1405369`

BIBLIOGRAPHY

[1] Y. E. Kim, M. S. Hipp, A. Bracher, M. Hayer-Hartl, and F. Ulrich Hartl, "Molecular chaperone functions in protein folding and proteostasis," *Annual review of biochemistry*, vol. 82, pp. 323–355, 2013.

[2] F. Kaiser, S. Bittrich, S. Salentin, C. Leberecht, V. J. Haupt, S. Krautwurst, M. Schroeder, and D. Labudde, "Backbone Brackets and Arginine Tweezers delineate Class I and Class II aminoacyl tRNA synthetases," *PLoS computational biology*, vol. 14, no. 4, p. e1006101, 2018.

[3] S. Ohno, *Evolution by Gene Duplication*. Springer-Verlag, 1970.

[4] R. J. Najmanovich, "Evolutionary studies of ligand binding sites in proteins," *Current opinion in structural biology*, vol. 45, pp. 85–90, 2017.

[5] A. Babajide, I. L. Hofacker, M. J. Sippl, and P. F. Stadler, "Neutral networks in protein space: A computational study based on knowledge-based potentials of mean force," *Folding and Design*, vol. 2, no. 5, pp. 261–269, 1997.

[6] G. C. Conant and K. H. Wolfe, "Turning a hobby into a job: how duplicated genes find new functions," *Nature Reviews Genetics*, vol. 9, no. 12, p. 938, 2008.

[7] R. E. Valas, S. Yang, and P. E. Bourne, "Nothing about protein structure classification makes sense except in the light of evolution," *Current opinion in structural biology*, vol. 19, no. 3, pp. 329–334, 2009.

[8] R. Pancsa, M. Varadi, P. Tompa, and W. F. Vranken, "Start2Fold: a database of hydrogen/deuterium exchange data on protein folding and stability," *Nucleic Acids Res.*, vol. 44, pp. D429–434, Jan 2016.

[9] R. L. Baldwin and G. D. Rose, "Is protein folding hierarchic? i. local structure and peptide folding," *Trends in biochemical sciences*, vol. 24, no. 1, pp. 26–33, 1999.

[10] R. L. Baldwin and G. D. Rose, "Is protein folding hierarchic? ii. folding intermediates and transition states," *Trends in biochemical sciences*, vol. 24, no. 2, pp. 77–83, 1999.

[11] D. Raimondi, G. Orlando, R. Pancsa, T. Khan, and W. F. Vranken, "Exploring the Sequence-based Prediction of Folding Initiation Sites in Proteins," *Sci Rep*, vol. 7, p. 8826, Aug 2017.

[12] S. W. Englander, L. Mayne, Z.-Y. Kan, and W. Hu, "Protein foldinghow and why: By hydrogen exchange, fragment separation, and mass spectrometry," *Annual review of biophysics*, vol. 45, pp. 135–152, 2016.

[13] R. Pancsa, D. Raimondi, E. Cilia, and W. F. Vranken, "Early Folding Events, Local Interactions, and Conservation of Protein Backbone Rigidity," *Biophys. J.*, vol. 110, pp. 572–583, Feb 2016.

[14] Y. Chen, F. Ding, and N. V. Dokholyan, "Fidelity of the protein structure reconstruction from inter-residue proximity constraints," *The Journal of Physical Chemistry B*, vol. 111, no. 25, pp. 7432–7438, 2007.

[15] R. Sathyapriya, J. M. Duarte, H. Stehr, I. Filippis, and M. Lappe, "Defining an essence of structure determining residue contacts in proteins," *PLoS computational biology*, vol. 5, no. 12, p. e1000584, 2009.

[16] J. M. Duarte, R. Sathyapriya, H. Stehr, I. Filippis, and M. Lappe, "Optimal contact definition for reconstruction of contact maps," *BMC bioinformatics*, vol. 11, no. 1, p. 283, 2010.

[17] M. Vassura, P. Di Lena, L. Margara, M. Mirto, G. Aloisio, P. Fariselli, and R. Casadio, "Blurring contact maps of thousands of proteins: what we can learn by reconstructing 3d structure," *BioData mining*, vol. 4, no. 1, p. 1, 2011.

[18] S. W. Englander and L. Mayne, "The nature of protein folding pathways," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 111, pp. 15873–15880, Nov 2014.

[19] S. W. Englander and L. Mayne, "The case for defined protein folding pathways," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 114, pp. 8253–8258, Aug 2017.

[20] E. Cilia, R. Pancsa, P. Tompa, T. Lenaerts, and W. F. Vranken, "From protein sequence to dynamics and disorder with DynaMine," *Nat Commun*, vol. 4, p. 2741, 2013.

[21] K. A. Dill, S. Bromberg, K. Yue, H. S. Chan, K. M. Ftebig, D. P. Yee, and P. D. Thomas, "Principles of protein foldinga perspective from simple exact models," *Protein science*, vol. 4, no. 4, pp. 561–602, 1995.

[22] O. B. Ptitsyn and K.-L. H. Ting, "Non-functional conserved residues in globins and their possible role as a folding nucleus," *Journal of molecular biology*, vol. 291, no. 3, pp. 671–682, 1999.

[23] R. P. Bhattacharyya, A. Remenyi, B. J. Yeh, and W. A. Lim, "Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits," *Annu. Rev. Biochem.*, vol. 75, pp. 655–680, 2006.

[24] M. Rorick, "Quantifying protein modularity and evolvability: a comparison of different techniques," *BioSystems*, vol. 110, pp. 22–33, Oct 2012.

[25] A. Gutteridge and J. M. Thornton, "Understanding nature's catalytic toolkit," *Trends Biochem. Sci.*, vol. 30, pp. 622–629, Nov 2005.

[26] A. A. Quadeer, D. Morales-Jimenez, and M. R. McKay, "Co-evolution networks of hiv/hcv are modular with direct association to structure and function," *PLOS Computational Biology*, vol. 14, pp. 1–29, 09 2018.

[27] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, "The protein folding problem," *Annu Rev Biophys*, vol. 37, pp. 289–316, 2008.

[28] E. Haglund, J. Danielsson, S. Kadhirvel, M. O. Lindberg, D. T. Logan, and M. Oliveberg, "Trimming down a protein structure to its bare foldons: spatial organization of the cooperative unit," *J. Biol. Chem.*, vol. 287, pp. 2731–2738, Jan 2012.

[29] L. A. Mirny and E. I. Shakhnovich, "Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function1," *Journal of molecular biology*, vol. 291, no. 1, pp. 177–196, 1999.

[30] C. B. Anfinsen, E. Haber, M. Sela, and F. White, "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain," *Proceedings of the National Academy of Sciences*, vol. 47, no. 9, pp. 1309–1314, 1961.

[31] C. Levinthal, "How to fold graciously," *Mossbauer spectroscopy in biological systems*, vol. 67, pp. 22–24, 1969.

[32] A. V. Finkelstein, A. J. Badretdin, O. V. Galzitskaya, D. N. Ivankov, N. S. Bogatyreva, and S. O. Garbuzynskiy, "There and back again: Two views on the protein folding puzzle," *Physics of life reviews*, vol. 21, pp. 56–71, 2017.

[33] R. Zwanzig, A. Szabo, and B. Bagchi, "Levinthal's paradox," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 89, pp. 20–22, Jan 1992.

[34] A. R. Fersht, "Nucleation mechanisms in protein folding," *Current opinion in structural biology*, vol. 7, no. 1, pp. 3–9, 1997.

[35] S. F. Betz, "Disulfide bonds and the stability of globular proteins," *Protein Science*, vol. 2, no. 10, pp. 1551–1558, 1993.

[36] S. Salentin, V. J. Haupt, S. Daminelli, and M. Schroeder, "Polypharmacology rescored: Protein–ligand interaction profiles for remote binding site similarity assessment," *Progress in biophysics and molecular biology*, vol. 116, no. 2-3, pp. 174–186, 2014.

[37] S. Salentin, S. Schreiber, V. J. Haupt, M. F. Adasme, and M. Schroeder, "PLIP: fully automated protein-ligand interaction profiler," *Nucleic Acids Res.*, vol. 43, pp. W443–447, Jul 2015.

[38] A. Razvi and J. M. Scholtz, "Lessons in stability from thermophilic proteins," *Protein Science*, vol. 15, no. 7, pp. 1569–1578, 2006.

[39] V. Frappier and R. Najmanovich, "Vibrational entropy differences between mesophile and thermophile proteins and their use in protein engineering," *Protein Science*, vol. 24, no. 4, pp. 474–483, 2015.

[40] D. F. Sticke, L. G. Presta, K. A. Dill, and G. D. Rose, "Hydrogen bonding in globular proteins," *Journal of molecular biology*, vol. 226, no. 4, pp. 1143–1159, 1992.

[41] C. N. Pace, H. Fu, K. Fryar, J. Landua, S. R. Trevino, D. Schell, R. L. Thurlkill, S. Imura, J. M. Scholtz, K. Gajiwala, *et al.*, "Contribution of hydrogen bonds to protein stability," *Protein Science*, vol. 23, no. 5, pp. 652–661, 2014.

[42] S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, and M. H. Hecht, "Protein design by binary patterning of polar and nonpolar amino acids," *Science*, vol. 262, no. 5140, pp. 1680–1685, 1993.

[43] H. S. Chan and K. A. Dill, "Transition states and folding dynamics of proteins and heteropolymers," *The Journal of chemical physics*, vol. 100, no. 12, pp. 9238–9257, 1994.

[44] T. Steiner, "The whole palette of hydrogen bonds," *Angew. Chem. Int. Ed*, vol. 41, no. 48-76, pp. 14–76, 2002.

[45] R. E. Hubbard and M. Kamran Haider, "Hydrogen bonds in proteins: role and strength," *eLS*, 2010.

[46] J. Gao, D. A. Bosco, E. T. Powers, and J. W. Kelly, "Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins," *Nature structural & molecular biology*, vol. 16, no. 7, p. 684, 2009.

[47] C. N. Pace, "Energetics of protein hydrogen bonds," *Nature Structural and Molecular Biology*, vol. 16, no. 7, p. 681, 2009.

[48] D. V. Waterhous and W. C. Johnson Jr, "Importance of environment in determining secondary structure in proteins," *Biochemistry*, vol. 33, no. 8, pp. 2121–2128, 1994.

[49] K. Shiraki, K. Nishikawa, and Y. Goto, "Trifluoroethanol-induced stabilization of the α-helical structure of β-lactoglobulin: implication for non-hierarchical protein folding," *Journal of molecular biology*, vol. 245, no. 2, pp. 180–194, 1995.

[50] S. Horowitz and R. C. Trievel, "Carbon-oxygen hydrogen bonding in biological structure and function," *Journal of Biological Chemistry*, vol. 287, no. 50, pp. 41576–41582, 2012.

[51] M. J. Plevin, D. L. Bryce, and J. Boisbouvier, "Direct detection of ch/π interactions in proteins," *Nature chemistry*, vol. 2, no. 6, pp. 466–471, 2010.

[52] C. N. Pace, H. Fu, K. L. Fryar, J. Landua, S. R. Trevino, B. A. Shirley, M. M. Hendricks, S. Iimura, K. Gajiwala, J. M. Scholtz, *et al.*, "Contribution of hydrophobic interactions to protein stability," *Journal of molecular biology*, vol. 408, no. 3, pp. 514–528, 2011.

[53] P. L. Privalov and S. J. Gill, "Stability of protein structure and hydrophobic interaction," in *Advances in protein chemistry*, vol. 39, pp. 191–234, Elsevier, 1988.

[54] K. Müller-Dethlefs and P. Hobza, "Noncovalent interactions: a challenge for experiment and theory," *Chemical Reviews*, vol. 100, no. 1, pp. 143–168, 2000.

[55] H. J. Dyson and P. E. Wright, "How does your protein fold? elucidating the apomyoglobin folding pathway," *Accounts of chemical research*, vol. 50, no. 1, pp. 105–111, 2016.

[56] H. Wirtz, S. Schäfer, C. Hoberg, K. M. Reid, D. M. Leitner, and M. Havenith, "Hydrophobic collapse of ubiquitin generates rapid protein-water motions," *Biochemistry*, 2018.

[57] H. Taketomi, Y. Ueda, and N. Gō, "Studies on protein folding, unfolding and fluctuations by computer simulation: I. the effect of specific amino acid sequence represented by specific inter-unit interactions," *International journal of peptide and protein research*, vol. 7, no. 6, pp. 445–459, 1975.

[58] K. F. Lau and K. A. Dill, "A lattice statistical mechanics model of the conformational and sequence spaces of proteins," *Macromolecules*, vol. 22, no. 10, pp. 3986–3997, 1989.

[59] S. Burley and G. A. Petsko, "Aromatic-aromatic interaction: a mechanism of protein structure stabilization," *Science*, vol. 229, no. 4708, pp. 23–28, 1985.

[60] E. Cauët, M. Rooman, R. Wintjens, J. Liévin, and C. Biot, "Histidine- aromatic interactions in proteins and protein- ligand complexes: quantum chemical study of x-ray and model structures," *Journal of chemical theory and computation*, vol. 1, no. 3, pp. 472–483, 2005.

[61] D. A. Dougherty, "Cation-π interactions involving aromatic amino acids," *The Journal of nutrition*, vol. 137, no. 6, pp. 1504S–1508S, 2007.

[62] J. P. Gallivan and D. A. Dougherty, "Cation-π interactions in structural biology," *Proceedings of the National Academy of Sciences*, vol. 96, no. 17, pp. 9459–9464, 1999.

[63] C. Biot, E. Buisine, J.-M. Kwasigroch, R. Wintjens, and M. Rooman, "Probing the energetic and structural role of amino acid/nucleobase cation-π interactions in protein-ligand complexes," *Journal of Biological Chemistry*, vol. 277, no. 43, pp. 40816–40822, 2002.

[64] D. J. Barlow and J. Thornton, "Ion-pairs in proteins," *Journal of molecular biology*, vol. 168, no. 4, pp. 867–885, 1983.

[65] O. R. Veltman, G. Vriend, F. Hardy, J. Mansfeld, B. Burg, G. Venema, and V. G. Eijsink, "Mutational analysis of a surface area that is critical for the thermal stability of thermolysin-like proteases," *The FEBS Journal*, vol. 248, no. 2, pp. 433–440, 1997.

[66] Z. S. Hendsch and B. Tidor, "Do salt bridges stabilize proteins? a continuum electrostatic analysis," *Protein Science*, vol. 3, no. 2, pp. 211–226, 1994.

[67] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips, "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis," *Nature*, vol. 181, pp. 662–666, Mar 1958.

[68] L. Pauling and R. B. Corey, "Atomic coordinates and structure factors for two helical configurations of polypeptide chains," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 37, pp. 235–240, May 1951.

[69] L. Pauling, R. B. Corey, and H. R. Branson, "The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 37, pp. 205–211, Apr 1951.

[70] A. A. Nickson, B. G. Wensley, and J. Clarke, "Take home lessons from studies of related proteins," *Current opinion in structural biology*, vol. 23, no. 1, pp. 66–74, 2013.

[71] V. Daggett and A. R. Fersht, "Is there a unifying mechanism for protein folding?," *Trends in biochemical sciences*, vol. 28, no. 1, pp. 18–25, 2003.

[72] R. R. Goluguri and J. B. Udgaonkar, "Microsecond rearrangements of hydrophobic clusters in an initially collapsed globule prime structure formation during the folding of a small protein," *Journal of molecular biology*, vol. 428, no. 15, pp. 3102–3117, 2016.

[73] R. L. Baldwin, "The search for folding intermediates and the mechanism of protein folding," *Annu. Rev. Biophys.*, vol. 37, pp. 1–21, 2008.

[74] T. R. Sosnick and D. Barrick, "The folding of single domain proteinshave we reached a consensus?," *Current opinion in structural biology*, vol. 21, no. 1, pp. 12–24, 2011.

[75] M. Karplus and D. L. Weaver, "Protein folding dynamics: The diffusion-collision model and experimental data," *Protein Science*, vol. 3, no. 4, pp. 650–668, 1994.

[76] A. R. Fersht, "Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications," *Proceedings of the National Academy of Sciences*, vol. 92, no. 24, pp. 10869–10873, 1995.

[77] D. A. Debe, M. J. Carlson, and W. A. Goddard, "The topomer-sampling model of protein folding," *Proceedings of the National Academy of Sciences*, vol. 96, no. 6, pp. 2596–2601, 1999.

[78] M. Karplus and D. L. Weaver, "Diffusion–collision model for protein folding," *Biopolymers: Original Research on Biomolecules*, vol. 18, no. 6, pp. 1421–1437, 1979.

[79] O. Ptitsyn and A. Rashin, "A model of myoglobin self-organization," *Biophysical chemistry*, vol. 3, no. 1, pp. 1–20, 1975.

[80] J. A. McCammon, B. R. Gelin, and M. Karplus, "Dynamics of folded proteins," *Nature*, vol. 267, no. 5612, p. 585, 1977.

[81] A. M. Lesk and G. D. Rose, "Folding units in globular proteins," *Proceedings of the National Academy of Sciences*, vol. 78, no. 7, pp. 4304–4308, 1981.

[82] H. Maity, M. Maity, M. M. Krishna, L. Mayne, and S. W. Englander, "Protein folding: the stepwise assembly of foldon units," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, pp. 4741–4746, Mar 2005.

[83] M. J. Rooman, J. P. Kocher, and S. J. Wodak, "Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions," *Biochemistry*, vol. 31, pp. 10226–10238, Oct 1992.

[84] M. J. Rooman and S. J. Wodak, "Extracting information on folding from the amino acid sequence: consensus regions with preferred conformation in homologous proteins," *Biochemistry*, vol. 31, pp. 10239–10249, Oct 1992.

[85] K. A. Dill and H. S. Chan, "From levinthal to pathways to funnels," *Nature Structural & Molecular Biology*, vol. 4, no. 1, pp. 10–19, 1997.

[86] C. M. Deane, M. Dong, F. P. Huard, B. K. Lance, and G. R. Wood, "Cotranslational protein folding–fact or fiction?," *Bioinformatics*, vol. 23, pp. i142–148, Jul 2007.

[87] R. Li and C. Woodward, "The hydrogen exchange core and protein folding," *Protein Science*, vol. 8, no. 8, pp. 1571–1590, 1999.

[88] P. E. Leopold, M. Montal, and J. N. Onuchic, "Protein folding funnels: a kinetic approach to the sequence-structure relationship.," *Proceedings of the National Academy of Sciences*, vol. 89, no. 18, pp. 8721–8725, 1992.

[89] P. G. Wolynes, J. N. Onuchic, D. Thirumalai, *et al.*, "Navigating the folding routes," *SCIENCE-NEW YORK THEN WASHINGTON-*, pp. 1619–1619, 1995.

[90] R. L. Baldwin, "Clash between energy landscape theory and foldon-dependent protein folding," *Proceedings of the National Academy of Sciences*, vol. 114, no. 32, pp. 8442–8443, 2017.

[91] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, "Atomic-level description of ubiquitin folding," *Proceedings of the National Academy of Sciences*, p. 201218321, 2013.

[92] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How fast-folding proteins fold," *Science*, vol. 334, no. 6055, pp. 517–520, 2011.

[93] S. Wallin and H. S. Chan, "A critical assessment of the topomer search model of protein folding using a continuum explicit-chain model with extensive conformational sampling," *Protein science*, vol. 14, no. 6, pp. 1643–1660, 2005.

[94] M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus, "Three key residues form a critical contact network in a protein folding transition state," *Nature*, vol. 409, no. 6820, pp. 641–645, 2001.

[95] N. V. Dokholyan, L. Li, F. Ding, and E. I. Shakhnovich, "Topological determinants of protein folding," *Proceedings of the National Academy of Sciences*, vol. 99, no. 13, pp. 8637–8641, 2002.

[96] S. W. Englander, L. Mayne, and M. M. Krishna, "Protein folding and misfolding: mechanism and principles," *Q. Rev. Biophys.*, vol. 40, pp. 287–326, Nov 2007.

[97] C. R. Matthews, "Pathways of protein folding," *Annual review of biochemistry*, vol. 62, no. 1, pp. 653–683, 1993.

[98] S. L. Mayo and R. L. Baldwin, "Guanidinium chloride induction of partial unfolding in amide proton exchange in rnase a," *Science*, vol. 262, no. 5135, pp. 873–876, 1993.

[99] Y. Bai, J. S. Milne, L. Mayne, and S. W. Englander, "Primary structure effects on peptide group hydrogen exchange," *Proteins: Structure, Function, and Bioinformatics*, vol. 17, no. 1, pp. 75–86, 1993.

[100] G. P. Connelly, Y. Bai, M.-F. Jeng, and S. W. Englander, "Isotope effects in peptide group hydrogen exchange," *Proteins: Structure, Function, and Bioinformatics*, vol. 17, no. 1, pp. 87–92, 1993.

[101] R. Molday, S. Englander, and R. Kallen, "Primary structure effects on peptide group hydrogen exchange," *Biochemistry*, vol. 11, no. 2, pp. 150–158, 1972.

[102] R. L. Baldwin, "The nature of protein folding pathways: the classical versus the new view," *Journal of biomolecular NMR*, vol. 5, no. 2, pp. 103–109, 1995.

[103] L. E. Rosen, S. V. Kathuria, C. R. Matthews, O. Bilsel, and S. Marqusee, "Non-native structure appears in microseconds during the folding of e. coli rnase h," *Journal of molecular biology*, vol. 427, no. 2, pp. 443–453, 2015.

[104] A. Labhardt, "[7] folding intermediates studied by circular dichroism," in *Methods in enzymology*, vol. 131, pp. 126–135, Elsevier, 1986.

[105] Y. Bai, T. R. Sosnick, L. Mayne, and S. W. Englander, "Protein folding intermediates: native-state hydrogen exchange," *Science*, vol. 269, pp. 192–197, Jul 1995.

[106] A. R. Fersht and S. Sato, "Phi-value analysis and the nature of protein-folding transition states," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, pp. 7976–7981, May 2004.

[107] M. Oliveberg and P. G. Wolynes, "The experimental survey of protein-folding energy landscapes," *Q. Rev. Biophys.*, vol. 38, pp. 245–288, Aug 2005.

[108] C. A. Royer, "Probing protein folding and conformational transitions with fluorescence," *Chemical reviews*, vol. 106, no. 5, pp. 1769–1784, 2006.

[109] A. Vallée-Bélisle and S. W. Michnick, "Visualizing transient protein-folding intermediates by tryptophan-scanning mutagenesis," *Nature Structural and Molecular Biology*, vol. 19, no. 7, p. 731, 2012.

[110] W. C. Johnson, "Protein secondary structure and circular dichroism: a practical guide," *Proteins: Structure, Function, and Bioinformatics*, vol. 7, no. 3, pp. 205–214, 1990.

[111] C. Nishimura, S. Prytulla, H. J. Dyson, and P. E. Wright, "Conservation of folding pathways in evolutionarily distant globin sequences," *Nature Structural & Molecular Biology*, vol. 7, no. 8, pp. 679–686, 2000.

[112] L. Bartoli, E. Capriotti, P. Fariselli, P. L. Martelli, and R. Casadio, "The pros and cons of predicting protein contact maps," in *Protein Structure Prediction*, pp. 199–217, Springer, 2008.

[113] M. Vendruscolo, E. Kussell, and E. Domany, "Recovery of protein structure from contact maps," *Folding and Design*, vol. 2, no. 5, pp. 295–306, 1997.

[114] B. Adhikari, D. Bhattacharya, R. Cao, and J. Cheng, "Confold: residue-residue contact-guided ab initio protein folding," *Proteins: Structure, Function, and Bioinformatics*, vol. 83, no. 8, pp. 1436–1449, 2015.

[115] B. M. Konopka, M. Ciombor, M. Kurczynska, and M. Kotulska, "Automated procedure for contact-map-based protein structure reconstruction," *The Journal of membrane biology*, vol. 247, no. 5, pp. 409–420, 2014.

[116] M. Vassura, L. Margara, P. Di Lena, F. Medri, P. Fariselli, and R. Casadio, "Ft-comar: fault tolerant three-dimensional structure reconstruction from protein contact maps," *Bioinformatics*, vol. 24, no. 10, pp. 1313–1315, 2008.

[117] B. Adhikari, J. Nowotny, D. Bhattacharya, J. Hou, and J. Cheng, "Coneva: a toolbox for comprehensive assessment of protein contacts," *BMC bioinformatics*, vol. 17, no. 1, p. 517, 2016.

[118] C. Vehlow, H. Stehr, M. Winkelmann, J. M. Duarte, L. Petzold, J. Dinse, and M. Lappe, "Cmview: interactive contact map visualization and analysis," *Bioinformatics*, vol. 27, no. 11, pp. 1573–1574, 2011.

[119] M. Kayikci, A. Venkatakrishnan, J. Scott-Brown, C. N. Ravarani, T. Flock, and M. M. Babu, "Visualization and analysis of non-covalent contacts using the protein contacts atlas," tech. rep., Nature Publishing Group, 2018.

[120] G. Faure, A. Bornot, and A. G. de Brevern, "Protein contacts, inter-residue interactions and side-chain modelling," *Biochimie*, vol. 90, pp. 626–639, Apr 2008.

[121] I. Anishchenko, S. Ovchinnikov, H. Kamisetty, and D. Baker, "Origins of coevolution between residues distant in protein 3D structures," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 114, pp. 9122–9127, Aug 2017.

[122] M. Vassura, L. Margara, P. Di Lena, F. Medri, P. Fariselli, and R. Casadio, "Reconstruction of 3d structures from protein contact maps," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 5, no. 3, pp. 357–367, 2008.

[123] B. Adhikari and J. Cheng, "Improved protein structure reconstruction using secondary structures, contacts at higher distance thresholds, and non-contacts," *BMC bioinformatics*, vol. 18, no. 1, p. 380, 2017.

[124] F. Simkovic, S. Ovchinnikov, D. Baker, and D. J. Rigden, "Applications of contact predictions to structural biology," *IUCrJ*, vol. 4, pp. 291–300, May 2017.

[125] D. Mercadante, F. Gräter, and C. Daday, "Conan: A tool to decode dynamical information from molecular interaction maps," *Biophysical journal*, vol. 114, no. 6, pp. 1267–1273, 2018.

[126] A. Raval, S. Piana, M. P. Eastwood, and D. E. Shaw, "Assessment of the utility of contact-based restraints in accelerating the prediction of protein structure using molecular dynamics simulations," *Protein Science*, vol. 25, no. 1, pp. 19–29, 2016.

[127] S. H. de Oliveira, E. C. Law, J. Shi, and C. M. Deane, "Sequential search leads to faster, more efficient fragment-based de novo protein structure prediction.," *Bioinformatics*, vol. 1, p. 9, 2017.

[128] M. Vendruscolo and E. Domany, "Protein folding using contact maps.," *Vitamins and hormones*, vol. 58, pp. 171–212, 2000.

[129] A. Caprara, R. Carr, S. Istrail, G. Lancia, and B. Walenz, "1001 optimal pdb structure alignments: integer programming methods for finding the maximum contact map overlap," *Journal of Computational Biology*, vol. 11, no. 1, pp. 27–52, 2004.

[130] R. Melo, C. Ribeiro, C. Murray, C. Veloso, C. da Silveira, G. Neshich, W. Meira Jr, R. Carceroni, and M. Santoro, "Finding protein-protein interaction patterns by contact map matching," *Genet. Mol. Res*, vol. 6, no. 4, pp. 946–963, 2007.

[131] T. F. Havel, "Distance geometry: Theory, algorithms, and chemical applications," *Encyclopedia of Computational Chemistry*, vol. 120, pp. 723–742, 1998.

[132] A. T. Brünger, P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J.-S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, *et al.*, "Crystallography & nmr system: a new software suite for macromolecular structure determination," *Acta Crystallographica Section D: Biological Crystallography*, vol. 54, no. 5, pp. 905–921, 1998.

[133] J. W. Ponder *et al.*, "Tinker: Software tools for molecular design," *Washington University School of Medicine, Saint Louis, MO*, vol. 3, 2004.

[134] N. Eswar, B. Webb, M. A. Marti-Renom, M. Madhusudhan, D. Eramian, M.-y. Shen, U. Pieper, and A. Sali, "Comparative protein structure modeling using modeller," *Current protocols in bioinformatics*, vol. 15, no. 1, pp. 5–6, 2006.

[135] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, "Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model," *PLoS Comput. Biol.*, vol. 13, p. e1005324, Jan 2017.

[136] Y. Zhang and J. Skolnick, "Tm-align: a protein structure alignment algorithm based on the tm-score," *Nucleic acids research*, vol. 33, no. 7, pp. 2302–2309, 2005.

[137] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins*, vol. 57, pp. 702–710, Dec 2004.

[138] J. Xu and Y. Zhang, "How significant is a protein structure similarity with TM-score = 0.5?," *Bioinformatics*, vol. 26, pp. 889–895, Apr 2010.

[139] S. de Oliveira and C. Deane, "Co-evolution techniques are reshaping the way we do structural bioinformatics," *F1000Research*, vol. 6, pp. 1–6, 2017.

[140] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, "Protein 3D structure computed from evolutionary sequence variation," *PLoS ONE*, vol. 6, no. 12, p. e28766, 2011.

[141] D. S. Marks, T. A. Hopf, and C. Sander, "Protein structure prediction from sequence variation," *Nat. Biotechnol.*, vol. 30, pp. 1072–1080, Nov 2012.

[142] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks, "Three-dimensional structures of membrane proteins from genomic sequencing," *Cell*, vol. 149, pp. 1607–1621, Jun 2012.

[143] T. A. Hopf, C. P. Scharfe, J. P. Rodrigues, A. G. Green, O. Kohlbacher, C. Sander, A. M. Bonvin, and D. S. Marks, "Sequence co-evolution gives 3D contacts and structures of protein complexes," *Elife*, vol. 3, Sep 2014.

[144] T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. Scharfe, M. Springer, C. Sander, and D. S. Marks, "Mutation effects predicted from sequence co-variation," *Nat. Biotechnol.*, vol. 35, pp. 128–135, Feb 2017.

[145] T. Liu, G. W. Tang, and E. Capriotti, "Comparative modeling: The state of the art and protein drug target structure prediction," *Combinatorial Chemistry & High Throughput Screening*, vol. 14, pp. 532–547, jul 2011.

[146] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, "Specific nucleus as the transition state for protein folding: evidence from the lattice model," *Biochemistry*, vol. 33, no. 33, pp. 10026–10036, 1994.

[147] C. Vieille and G. J. Zeikus, "Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability," *Microbiology and molecular biology reviews*, vol. 65, no. 1, pp. 1–43, 2001.

[148] P. Wozniak, B. Konopka, J. Xu, G. Vriend, and M. Kotulska, "Forecasting residue–residue contact prediction accuracy," *Bioinformatics*, vol. 33, no. 21, pp. 3405–3414, 2017.

[149] L. Oliveira, P. B. Paiva, A. Paiva, and G. Vriend, "Identification of functionally conserved residues with the use of entropy–variability plots," *Proteins: Structure, Function, and Bioinformatics*, vol. 52, no. 4, pp. 544–552, 2003.

[150] D. T. Jones, D. W. Buchan, D. Cozzetto, and M. Pontil, "Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments," *Bioinformatics*, vol. 28, no. 2, pp. 184–190, 2011.

[151] S. Seemayer, M. Gruber, and J. Söding, "Ccmpredfast and precise prediction of protein residue–residue contacts from correlated mutations," *Bioinformatics*, vol. 30, no. 21, pp. 3128–3130, 2014.

[152] H. Kamisetty, S. Ovchinnikov, and D. Baker, "Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 15674–15679, sep 2013.

[153] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, "Improved contact prediction in proteins: using pseudolikelihoods to infer potts models," *Physical Review E*, vol. 87, no. 1, p. 012707, 2013.

[154] S. Wu and Y. Zhang, "A comprehensive assessment of sequence-based and template-based methods for protein contact prediction," *Bioinformatics*, vol. 24, no. 7, pp. 924–931, 2008.

[155] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[156] P. Di Lena, K. Nagata, and P. Baldi, "Deep architectures for protein contact map prediction," *Bioinformatics*, vol. 28, no. 19, pp. 2449–2457, 2012.

[157] M. J. Skwark, D. Raimondi, M. Michel, and A. Elofsson, "Improved contact predictions using the recognition of protein like contact patterns," *PLoS computational biology*, vol. 10, no. 11, p. e1003889, 2014.

[158] D. T. Jones, T. Singh, T. Kosciolek, and S. Tetchner, "Metapsicov: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins," *Bioinformatics*, vol. 31, no. 7, pp. 999–1006, 2014.

[159] Z. Wang and J. Xu, "Predicting protein contact map using evolutionary and physical constraints by integer programming," *Bioinformatics*, vol. 29, no. 13, pp. i266–i273, 2013.

[160] J. Ma, S. Wang, Z. Wang, and J. Xu, "Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning," *Bioinformatics*, vol. 31, no. 21, pp. 3506–3513, 2015.

[161] B. B. Kragelund, J. Knudsen, and F. M. Poulsen, "Local perturbations by ligand binding of hydrogen deuterium exchange kinetics in a four-helix bundle protein, acyl coenzyme a binding protein (acbp)," *Journal of molecular biology*, vol. 250, no. 5, pp. 695–706, 1995.

[162] C. Merstorf, O. Maciejak, J. Mathe, M. Pastoriza-Gallego, B. Thiebot, M.-J. Clement, J. Pelta, L. Auvray, P. A. Curmi, and P. Savarin, "Mapping the conformational stability of maltose binding protein at the residue scale using nuclear magnetic resonance hydrogen exchange experiments," *Biochemistry*, vol. 51, no. 44, pp. 8919–8930, 2012.

[163] M. M. Krishna, L. Hoang, Y. Lin, and S. W. Englander, "Hydrogen exchange methods to study protein folding," *Methods*, vol. 34, no. 1, pp. 51–64, 2004.

[164] R. C. Holland, T. A. Down, M. Pocock, A. Prlić, D. Huen, K. James, S. Foisy, A. Dräger, A. Yates, M. Heuer, *et al.*, "Biojava: an open-source framework for bioinformatics," *Bioinformatics*, vol. 24, no. 18, pp. 2096–2097, 2008.

[165] A. Prlić, A. Yates, S. E. Bliven, P. W. Rose, J. Jacobsen, P. V. Troshin, M. Chapman, J. Gao, C. H. Koh, S. Foisy, *et al.*, "Biojava: an open-source framework for bioinformatics in 2012," *Bioinformatics*, vol. 28, no. 20, pp. 2693–2695, 2012.

[166] A. Shrake and J. Rupley, "Environment and exposure to solvent of protein atoms. lysozyme and insulin," *Journal of molecular biology*, vol. 79, no. 2, pp. 351IN15365–364371, 1973.

[167] W. Kabsch and C. Sander, "Dssp: definition of secondary structure of proteins given a set of 3d coordinates," *Biopolymers*, vol. 22, pp. 2577–2637, 1983.

[168] B. Rost and C. Sander, "Conservation and prediction of solvent accessibility in protein families," *Proteins: Structure, Function, and Bioinformatics*, vol. 20, no. 3, pp. 216–226, 1994.

[169] B. Monastyrskyy, D. D'Andrea, K. Fidelis, A. Tramontano, and A. Kryshtafovych, "Evaluation of residue–residue contact prediction in casp10," *Proteins: Structure, Function, and Bioinformatics*, vol. 82, no. S2, pp. 138–153, 2014.

[170] E. Shakhnovich and A. Gutin, "Implications of thermodynamics of protein folding for evolution of primary sequences," *Nature*, vol. 346, no. 6286, p. 773, 1990.

[171] T. Alber, S. Dao-Pin, K. Wilson, J. A. Wozniak, S. P. Cook, and B. W. Matthews, "Contributions of hydrogen bonds of thr 157 to the thermodynamic stability of phage t4 lysozyme," *Nature*, vol. 330, no. 6143, p. 41, 1987.

[172] N. Tokuriki, F. Stricher, L. Serrano, and D. S. Tawfik, "How protein stability and new functions trade off," *PLoS Computational Biology*, vol. 4, no. 2, p. e1000002, 2008.

[173] U. Consortium *et al.*, "Uniprot: a hub for protein information," *Nucleic acids research*, p. gku989, 2014.

[174] S. Bittrich, M. Schroeder, and D. Labudde, "Characterizing the relation of functional and early folding residues in protein structures using the example of aminoacyl-trna synthetases," *PLOS ONE*, vol. 13, pp. 1–23, 10 2018.

[175] D. E. Kim, Q. Yi, S. T. Gladwin, J. M. Goldberg, and D. Baker, "The single helix in protein l is largely disrupted at the rate-limiting step in folding1," *Journal of molecular biology*, vol. 284, no. 3, pp. 807–815, 1998.

[176] A. R. Panchenko, Z. Luthey-Schulten, and P. G. Wolynes, "Foldons, protein structural modules, and exons," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 93, pp. 2008–2013, Mar 1996.

[177] A. Krishnan, A. Giuliani, J. P. Zbilut, and M. Tomita, "Network scaling invariants help to elucidate basic topological principles of proteins," *J. Proteome Res.*, vol. 6, pp. 3924–3934, Oct 2007.

[178] F. Sinibaldi, M. C. Piro, B. D. Howes, G. Smulevich, F. Ascoli, and R. Santucci, "Rupture of the hydrogen bond linking two ω-loops induces the molten globule state at neutral ph in cytochrome c," *Biochemistry*, vol. 42, no. 24, pp. 7604–7610, 2003.

[179] S. Zaidi, M. I. Hassan, A. Islam, and F. Ahmad, "The role of key residues in structure, function, and stability of cytochrome-c," *Cellular and molecular life sciences*, vol. 71, no. 2, pp. 229–255, 2014.

[180] H. Roder, G. A. Elove, and S. W. Englander, "Structural characterization of folding intermediates in cytochrome c by H-exchange labelling and proton NMR," *Nature*, vol. 335, pp. 700–704, Oct 1988.

[181] R. Aurora and G. D. Rosee, "Helix capping," *Protein Science*, vol. 7, no. 1, pp. 21–38, 1998.

[182] S. Bittrich, F. Heinke, and D. Labudde, "equant - a server for fast protein model quality assessment by integrating high-dimensional data and machine learning," *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery*, pp. 419–433, 2016.

[183] F. Heinke, S. Schildbach, D. Stockmann, and D. Labudde, "eprosa database and toolbox for investigating protein sequence–structure–function relationships through energy profiles," *Nucleic acids research*, vol. 41, no. D1, pp. D320–D326, 2012.

[184] M. Silow and M. Oliveberg, "Transient aggregates in protein folding are easily mistaken for folding intermediates," *Proceedings of the National Academy of Sciences*, vol. 94, no. 12, pp. 6084–6086, 1997.

[185] S. Velankar, J. M. Dana, J. Jacobsen, G. van Ginkel, P. J. Gane, J. Luo, T. J. Oldfield, C. ODonovan, M.-J. Martin, and G. J. Kleywegt, "Sifts: structure integration with function, taxonomy and sequences resource," *Nucleic acids research*, vol. 41, no. D1, pp. D483–D489, 2012.

[186] M. Vendruscolo, N. V. Dokholyan, E. Paci, and M. Karplus, "Small-world view of the amino acids that play a key role in protein folding," *Physical Review E*, vol. 65, no. 6, p. 061910, 2002.

[187] L. Royer, M. Reimann, B. Andreopoulos, and M. Schroeder, "Unraveling protein networks with power graph analysis," *PLoS computational biology*, vol. 4, no. 7, p. e1000108, 2008.

[188] M. M. Rorick and G. P. Wagner, "Protein structural modularity and robustness are associated with evolvability," *Genome biology and evolution*, vol. 3, pp. 456–475, 2011.

[189] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, pp. 35–41, 1977.

[190] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, and S. Pietrokovski, "Network analysis of protein structures identifies functional residues," *Journal of molecular biology*, vol. 344, no. 4, pp. 1135–1146, 2004.

[191] M. J. Sippl, "Boltzmann's principle, knowledge-based mean fields and protein folding. an approach to the computational determination of protein structures," *Journal of computer-aided molecular design*, vol. 7, no. 4, pp. 473–501, 1993.

[192] M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl, "Identification of native protein folds amongst a large number of incorrect models: the calculation of low energy conformations from potentials of mean force," *Journal of molecular biology*, vol. 216, no. 1, pp. 167–180, 1990.

[193] M. J. Sippl, "Calculation of conformational ensembles from potentials of mena force: an approach to the knowledge-based prediction of local structures in globular proteins," *Journal of molecular biology*, vol. 213, no. 4, pp. 859–883, 1990.

[194] M. J. Sippl, "Recognition of errors in three-dimensional structures of proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 17, no. 4, pp. 355–362, 1993.

[195] M. J. Sippl, "Knowledge-based potentials for proteins," *Current opinion in structural biology*, vol. 5, no. 2, pp. 229–235, 1995.

[196] G. Verkhivker, K. Appelt, S. Freer, and J. Villafranca, "Empirical free energy calculations of ligand-protein crystallographic complexes. i. knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity," *Protein Engineering, Design and Selection*, vol. 8, no. 7, pp. 677–691, 1995.

[197] D. T. Jones, W. Taylort, and J. M. Thornton, "A new approach to protein fold recognition," *Nature*, vol. 358, no. 6381, p. 86, 1992.

[198] P. Benkert, S. C. E. Tosatto, and D. Schomburg, "Qmean: A comprehensive scoring function for model quality assessment," *Proteins: Struct., Funct., Bioinf.*, vol. 71, pp. 261–277, Apr 2008.

[199] F. Melo, D. Devos, E. Depiereux, and E. Feytmans, "Anolea: a www server to assess protein structures.," in *ISMB*, vol. 5, pp. 187–190, 1997.

[200] R. Bryll, R. Gutierrez-Osuna, and F. Quek, "Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets," *Pattern recognition*, vol. 36, no. 6, pp. 1291–1302, 2003.

[201] G. Holmes, A. Donkin, and I. H. Witten, "Weka: A machine learning workbench," in *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, pp. 357–361, IEEE, 1994.

[202] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, "Data mining in bioinformatics using weka," *Bioinformatics*, vol. 20, pp. 2479–2481, apr 2004.

[203] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[204] P. Benkert, M. Künzli, and T. Schwede, "Qmean server for protein model quality estimation," *Nucleic acids research*, vol. 37, no. suppl_2, pp. W510–W514, 2009.

[205] J. M. Chambers, *Graphical Methods for Data Analysis: 0*. Chapman and Hall/CRC, 2017.

[206] W. F. DeGrado, H. Gratkowski, and J. D. Lear, "How do helix–helix interactions help determine the folds of membrane proteins? perspectives from the study of homo-oligomeric helical bundles," *Protein Science*, vol. 12, no. 4, pp. 647–665, 2003.

[207] W. Dyrka, J.-C. Nebel, and M. Kotulska, "Probabilistic grammatical model for helix-helix contact site classification.," *Algorithms for molecular biology: AMB*, vol. 8, no. 1, pp. 31–31, 2013.

[208] K. Brinda and S. Vishveshwara, "A network representation of protein structures: implications for protein stability," *Biophysical journal*, vol. 89, no. 6, pp. 4159–4170, 2005.

[209] D. Balchin, M. Hayer-Hartl, and F. U. Hartl, "In vivo aspects of protein folding and quality control," *Science*, vol. 353, no. 6294, p. aac4354, 2016.

[210] F. Cymer, G. von Heijne, and S. H. White, "Mechanisms of integral membrane protein insertion and folding," *Journal of molecular biology*, vol. 427, no. 5, pp. 999–1022, 2015.

[211] R. Guimera and L. A. Nunes Amaral, "Functional cartography of complex metabolic networks," *Nature*, vol. 433, pp. 895–900, Feb 2005.

[212] Y. Levy, "Protein assembly and building blocks: Beyond the limits of the lego brick metaphor," *Biochemistry*, vol. 56, no. 38, pp. 5040–5048, 2017.

[213] P. G. Wolynes, "Three paradoxes of protein folding," *Protein folds: A Distances Based Approach*, pp. 3–17, 1996.

[214] J. S. Hleap, E. Susko, and C. Blouin, "Defining structural and evolutionary modules in proteins: a community detection approach to explore sub-domain architecture," *BMC Struct. Biol.*, vol. 13, p. 20, Oct 2013.

[215] K. Teilum, B. B. Kragelund, J. Knudsen, and F. M. Poulsen, "Formation of hydrogen bonds precedes the rate-limiting formation of persistent structure in the folding of acbp," *Journal of molecular biology*, vol. 301, no. 5, pp. 1307–1314, 2000.

[216] B. B. Kragelund, K. V. Andersen, J. C. Madsen, J. Knudsen, and F. M. Poulsen, "Three-dimensional structure of the complex between acyl-coenzyme a binding protein and palmitoyl-coenzyme a," *Journal of molecular biology*, vol. 230, no. 4, pp. 1260–1277, 1993.

[217] R. Nussinov and B. Ma, "Protein dynamics and conformational selection in bidirectional signal transduction," *BMC biology*, vol. 10, no. 1, p. 2, 2012.

[218] L. A. Mirny, V. I. Abkevich, and E. I. Shakhnovich, "How evolution makes proteins fold quickly," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 95, pp. 4976–4981, Apr 1998.

[219] J. D. Bloom, S. T. Labthavikul, C. R. Otey, and F. H. Arnold, "Protein stability promotes evolvability," *Proceedings of the National Academy of Sciences*, vol. 103, no. 15, pp. 5869–5874, 2006.

[220] C. M. Dobson and M. Karplus, "The fundamentals of protein folding: bringing together theory and experiment," *Curr. Opin. Struct. Biol.*, vol. 9, pp. 92–101, Feb 1999.

[221] F. Kaiser, A. Eisold, and D. Labudde, "A Novel Algorithm for Enhanced Structural Motif Matching in Proteins," *J. Comput. Biol.*, vol. 22, pp. 698–713, Jul 2015.

[222] F. Kaiser and D. Labudde, "Unsupervised discovery of geometrically common structural motifs and long-range contacts in protein 3d structures," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2017.

[223] H. R. Brodkin, N. A. DeLateur, S. Somarowthu, C. L. Mills, W. R. Novak, P. J. Beuning, D. Ringe, and M. J. Ondrechen, "Prediction of distal residue participation in enzyme catalysis," *Protein Science*, vol. 24, no. 5, pp. 762–778, 2015.

[224] S. M. Larson, I. Ruczinski, A. R. Davidson, D. Baker, and K. W. Plaxco, "Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation1," *Journal of molecular biology*, vol. 316, no. 2, pp. 225–233, 2002.

[225] Y. Y. Tseng and J. Liang, "Are residues in a protein folding nucleus evolutionarily conserved?," *Journal of molecular biology*, vol. 335, no. 4, pp. 869–880, 2004.

[226] T. M. Jacobs and B. Kuhlman, "Using anchoring motifs for the computational design of protein-protein interactions," *Biochem. Soc. Trans.*, vol. 41, pp. 1141–1145, Oct 2013.

[227] M. Hecht, Y. Bromberg, and B. Rost, "News from the protein mutability landscape," *J. Mol. Biol.*, vol. 425, pp. 3937–3948, Nov 2013.

[228] B. Hammer and T. Villmann, "Generalized relevance learning vector quantization," *Neural Networks*, vol. 15, no. 8, pp. 1059–1068, 2002.

[229] M. Kästner, B. Hammer, M. Biehl, and T. Villmann, "Functional relevance learning in generalized learning vector quantization," *Neurocomputing*, vol. 90, pp. 85–95, aug 2012.

[230] F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: large-scale single-cell gene expression data analysis," *Genome Biol.*, vol. 19, p. 15, 02 2018.

[231] A. Sato and K. Yamada, "Generalized learning vector quantization," in *Advances in Neural Information Processing Systems 8* (D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, eds.), pp. 423–429, Cambridge, MA: MIT Press, 1996.

[232] P. Schneider, M. Biehl, and B. Hammer, "Distance learning in discriminative vector quantization," *Neural Computation*, vol. 21, pp. 2942–2969, oct 2009.

[233] C. B. Anfinsen and H. A. Scheraga, "Experimental and theoretical aspects of protein folding," *Adv. Protein Chem.*, vol. 29, pp. 205–300, 1975.

[234] T. Kohonen, "Learning vector quantization for pattern recognition," tech. rep., TKK-F-A601, Helsinki, 1986.

[235] M. Kaden, M. Lange, D. Nebel, M. Riedel, T. Geweniger, and T. Villmann, "Aspects in classification learning-review of recent developments in learning vector quantization," *Foundations of Computing and Decision Sciences*, vol. 39, no. 2, pp. 79–105, 2014.

[236] M. Kaden, *Integration of Auxiliary Data Knowledge in Prototype Based Vector Quantization and Classification Models*. PhD thesis, University Leipzig, Germany, 2015.

[237] T. Kohonen, "Learning vector quantization," in *Self-Organizing Maps*, pp. 203–217, Springer, 1997.

[238] K. Bunte, P. Schneider, B. Hammer, F. Schleif, T. Villmann, and M. Biehl, "Limited rank matrix learning, discriminative dimension reduction and visualization," *Neural Networks*, vol. 26, pp. 159–173, 2012.

[239] N. V. Chawla, *Data Mining and Knowledge Discovery Handbook*, ch. Data Mining for Imbalanced Datasets: An Overview, pp. 875–886. Boston, MA: Springer US, 2010.

[240] T. Fawcett, "An introduction to roc analysis," *Pattern Recogn. Lett.*, vol. 27, pp. 861–874, jun 2006.

[241] K. A. Dill, "Theory for the folding and stability of globular proteins," *Biochemistry*, vol. 24, no. 6, pp. 1501–1509, 1985.

[242] P. F. Faísca, "The nucleation mechanism of protein folding: a survey of computer simulation studies," *Journal of Physics: Condensed Matter*, vol. 21, no. 37, p. 373102, 2009.

[243] M. M. Gromiha, "Multiple contact network is a key determinant to protein folding rates," *Journal of chemical information and modeling*, vol. 49, no. 4, pp. 1130–1135, 2009.

[244] S. N. Rodin and S. Ohno, "Two types of aminoacyl-tRNA synthetases could be originally encoded by complementary strands of the same nucleic acid," *Orig Life Evol Biosph*, vol. 25, pp. 565–589, Dec 1995.

[245] M. Ibba and D. Söll, "Aminoacyl-tRNA synthesis," *Annu. Rev. Biochem.*, vol. 69, pp. 617–650, 2000.

[246] J. T. Wong, "Coevolution theory of the genetic code at age thirty," *Bioessays*, vol. 27, pp. 416–425, Apr 2005.

[247] L. Martinez-Rodriguez, O. Erdogan, M. Jimenez-Rodriguez, K. Gonzalez-Rivera, T. Williams, L. Li, V. Weinreb, M. Collier, S. N. Chandrasekaran, X. Ambroggio, *et al.*, "Functional class i and ii amino acid-activating enzymes can be coded by opposite strands of the same gene," *Journal of Biological Chemistry*, vol. 290, no. 32, pp. 19710–19725, 2015.

[248] C. W. Carter, "Coding of class i and ii aminoacyl-trna synthetases," in *Protein Reviews*, pp. 103–148, Springer, 2017.

[249] J. J. Burbaum and P. Schimmel, "Structural relationships and the classification of aminoacyl-tRNA synthetases," *J Biol Chem*, vol. 266, no. 26, pp. 16965–16968, 1991.

[250] L. R. de Pouplana and P. Schimmel, "Aminoacyl-trna synthetases: potential markers of genetic code development," *Trends in biochemical sciences*, vol. 26, no. 10, pp. 591–596, 2001.

[251] G. Eriani, M. Delarue, O. Poch, J. Gangloff, and D. Moras, "Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs," *Nature*, vol. 347, no. 6289, p. 203, 1990.

[252] Y. I. Wolf, L. Aravind, N. V. Grishin, and E. V. Koonin, "Evolution of aminoacyl-trna synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events," *Genome research*, vol. 9, no. 8, pp. 689–710, 1999.

[253] D. Moras, "Structural and functional relationships between aminoacyl-tRNA synthetases," *Trends Biochem. Sci.*, vol. 17, pp. 159–164, Apr 1992.

[254] C. W. Carter and R. Wolfenden, "trna acceptor stem and anticodon bases form independent codes related to protein folding," *Proceedings of the National Academy of Sciences*, vol. 112, no. 24, pp. 7489–7494, 2015.

[255] A. Chaliotis, P. Vlastaridis, D. Mossialos, M. Ibba, H. D. Becker, C. Stathopoulos, and G. D. Amoutzias, "The complex evolutionary history of aminoacyl-tRNA synthetases," *Nucleic Acids Res.*, vol. 45, pp. 1059–1068, 02 2017.

[256] C. D. Livingstone and G. J. Barton, "Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation," *Comput. Appl. Biosci.*, vol. 9, pp. 745–756, Dec 1993.

[257] H. Belrhali, A. Yaremchuk, M. Tukalo, C. Berthet-Colominas, B. Rasmussen, P. Bosecke, O. Diat, and S. Cusack, "The structural basis for seryl-adenylate and Ap4A synthesis by seryl-tRNA synthetase," *Structure*, vol. 3, pp. 341–352, Apr 1995.

[258] M. Fujinaga, C. Berthet-Colominas, A. D. Yaremchuk, M. A. Tukalo, and S. Cusack, "Refined crystal structure of the seryl-tRNA synthetase from Thermus thermophilus at 2.5 A resolution," *J. Mol. Biol.*, vol. 234, pp. 222–233, Nov 1993.

[259] A. Ambrogelly, D. Söll, O. Nureki, S. Yokoyama, and M. Ibba, *Class I Lysyl-tRNA Synthetases*. Landes Bioscience, 2013.

[260] E. Schmitt, M. Panvert, S. Blanquet, and Y. Mechulam, "Transition state stabilization by the highmotif of class i aminoacyl-trna synthetases: the case of escherichia coli methionyl-trna synthetase," *Nucleic acids research*, vol. 23, no. 23, pp. 4793–4798, 1995.

[261] G. Eriani, J. Cavarelli, F. Martin, G. Dirheimer, D. Moras, and J. Gangloff, "Role of dimerization in yeast aspartyl-trna synthetase and importance of the class ii invariant proline," *Proceedings of the National Academy of Sciences*, vol. 90, no. 22, pp. 10816–10820, 1993.

[262] A. Åberg, A. Yaremchuk, M. Tukalo, B. Rasmussen, and S. Cusack, "Crystal structure analysis of the activation of histidine by thermus thermophilus histidyl-trna synthetase," *Biochemistry*, vol. 36, no. 11, pp. 3084–3094, 1997.

[263] S. Cusack, "Aminoacyl-trna synthetases," *Current opinion in structural biology*, vol. 7, no. 6, pp. 881–889, 1997.

[264] P. O'Donoghue and Z. Luthey-Schulten, "On the evolution of structure in aminoacyl-tRNA synthetases," *Microbiology and Molecular Biology Reviews*, vol. 67, no. 4, pp. 550–573, 2003.

[265] E. A. First and A. R. Fersht, "Involvement of threonine 234 in catalysis of tyrosyl adenylate formation by tyrosyl-tRNA synthetase," *Biochemistry*, vol. 32, pp. 13644–13650, Dec 1993.

[266] V. Weinreb, L. Li, and C. W. Carter, "A master switch couples Mg2+-assisted catalysis to domain motion in B. stearothermophilus tryptophanyl-tRNA Synthetase," *Structure*, vol. 20, pp. 128–138, Jan 2012.

[267] V. Weinreb, L. Li, S. N. Chandrasekaran, P. Koehl, M. Delarue, and C. W. Carter, "Enhanced amino acid selection in fully evolved tryptophanyl-tRNA synthetase, relative to its urzyme, requires domain motion sensed by the D1 switch, a remote dynamic packing motif," *J. Biol. Chem.*, vol. 289, pp. 4367–4376, Feb 2014.

[268] J. G. Arnez and D. Moras, "Structural and functional considerations of the aminoacylation reaction," *Trends in biochemical sciences*, vol. 22, no. 6, pp. 211–216, 1997.

[269] S. Cammer and C. W. Carter Jr, "Six rossmannoid folds, including the class i aminoacyl-trna synthetases, share a partial core with the anti-codon-binding domain of a class ii aminoacyl-trna synthetase," *Bioinformatics*, vol. 26, no. 6, pp. 709–714, 2010.

[270] S. Cusack, C. Berthet-Colominas, M. Hartlein, N. Nassar, and R. Leberman, "A second class of synthetase structure revealed by x-ray analysis of escherichia coli seryl-tRNA synthetase at 2.5 a," *Nature*, vol. 347, no. 6290, p. 249, 1990.

[271] S. Cusack, M. Hartlein, and R. Leberman, "Sequence, structural and evolutionary relationships between class 2 aminoacyl-tRNA synthetases," *Nucleic Acids Res.*, vol. 19, pp. 3489–3498, Jul 1991.

[272] C. R. Woese, G. J. Olsen, M. Ibba, and D. Söll, "Aminoacyl-trna synthetases, the genetic code, and the evolutionary process," *Microbiology and Molecular Biology Reviews*, vol. 64, no. 1, pp. 202–236, 2000.

[273] S. D. Banik and N. Nandi, "Mechanism of the activation step of the aminoacylation reaction: a significant difference between class I and class II synthetases," *J. Biomol. Struct. Dyn.*, vol. 30, no. 6, pp. 701–715, 2012.

[274] Y. Diaz-Lazcoz, J. Aude, P. Nitschke, H. Chiapello, C. Landes-Devauchelle, and J. Risler, "Evolution of genes, evolution of species: the case of aminoacyl-tRNA synthetases.," *Molecular biology and evolution*, vol. 15, no. 11, pp. 1548–1561, 1998.

[275] M. Ruff, S. Krishnaswamy, M. Boeglin, A. Poterszman, A. Mitschler, A. Podjarny, B. Rees, J. Thierry, and D. Moras, "Class ii aminoacyl transfer rna synthetases: crystal structure of yeast aspartyl-trna synthetase complexed with trna (asp)," *Science*, vol. 252, no. 5013, pp. 1682–1689, 1991.

[276] H. B. LeJohn, L. E. Cameron, B. Yang, and S. L. Rennie, "Molecular characterization of an nad-specific glutamate dehydrogenase gene inducible by l-glutamine. antisense gene pair arrangement with l-glutamine-inducible heat shock 70-like protein gene.," *Journal of Biological Chemistry*, vol. 269, no. 6, pp. 4523–4531, 1994.

[277] C. W. Carter and W. L. Duax, "Did trna synthetase classes arise on opposite strands of the same gene?," *Molecular cell*, vol. 10, no. 4, pp. 705–708, 2002.

[278] C. W. Carter, L. Li, V. Weinreb, M. Collier, K. Gonzalez-Rivera, M. Jimenez-Rodriguez, O. Erdogan, B. Kuhlman, X. Ambroggio, T. Williams, and S. N. Chandrasekharan, "The Rodin-Ohno hypothesis that two enzyme superfamilies descended from one ancestral gene: an unlikely scenario for the origins of translation that will not be dismissed," *Biol. Direct*, vol. 9, p. 11, Jun 2014.

[279] A. R. Fersht, "Dissection of the structure and activity of the tyrosyl-trna synthetase by site-directed mutagenesis," *Biochemistry*, vol. 26, no. 25, pp. 8031–8037, 1987.

[280] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman, "The Pfam protein families database: towards a more sustainable future," *Nucleic Acids Res.*, vol. 44, pp. D279–285, Jan 2016.

[281] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, pp. 443–453, Mar 1970.

[282] F. Armougom, S. Moretti, O. Poirot, S. Audic, P. Dumas, B. Schaeli, V. Keduas, and C. Notredame, "Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee," *Nucleic Acids Res.*, vol. 34, pp. W604–608, Jul 2006.

[283] A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton, "Jalview Version 2—a multiple sequence alignment editor and analysis workbench," *Bioinformatics*, vol. 25, pp. 1189–1191, May 2009.

[284] J. Cavarelli, G. Eriani, B. Rees, M. Ruff, M. Boeglin, A. Mitschler, F. Martin, J. Gangloff, J. C. Thierry, and D. Moras, "The active site of yeast aspartyl-tRNA synthetase: structural and functional aspects of the aminoacylation reaction," *EMBO J.*, vol. 13, pp. 327–337, Jan 1994.

[285] W. W. Navarre, S. B. Zou, H. Roy, J. L. Xie, A. Savchenko, A. Singer, E. Edvokimova, L. R. Prost, R. Kumar, M. Ibba, and F. C. Fang, "PoxA, yjeK, and elongation factor P coordinately modulate virulence and drug resistance in Salmonella enterica," *Mol. Cell*, vol. 39, pp. 209–221, Jul 2010.

[286] R. Giege and M. Springer, "Aminoacyl-tRNA Synthetases in the Bacterial World," *EcoSal Plus*, vol. 7, May 2016.

[287] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner, "WebLogo: a sequence logo generator," *Genome Res.*, vol. 14, pp. 1188–1190, Jun 2004.

[288] A. M. Gallina, P. Bork, and D. Bordo, "Structural analysis of protein-ligand interactions: the binding of endogenous compounds and of synthetic drugs," *J. Mol. Recognit.*, vol. 27, pp. 65–72, Feb 2014.

[289] W. Hol, P. T. Van Duijnen, and H. Berendsen, "The α-helix dipole and the properties of proteins," *Nature*, vol. 273, no. 5662, p. 443, 1978.

[290] M. Kapustina, V. Weinreb, L. Li, B. Kuhlman, and C. W. Carter Jr, "A conformational transition state accompanies tryptophan activation by b. stearothermophilus tryptophanyl-trna synthetase," *Structure*, vol. 15, no. 10, pp. 1272–1284, 2007.

[291] V. Weinreb, L. Li, C. L. Campbell, L. S. Kaguni, and C. W. Carter Jr, "Mg2+-assisted catalysis by b. stearothermophilus trprs is promoted by allosteric effects," *Structure*, vol. 17, no. 7, pp. 952–964, 2009.

TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Computer Science Division of Bioinformatics, Biotechnology Center TU Dresden

STATEMENT OF AUTHORSHIP

TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Computer Science Division of Bioinformatics, Biotechnology Center TU Dresden

I hereby certify that I have authored this Dissertation entitled *Understanding the Structural and Functional Importance of Early Folding Residues in Protein Structures* independently and without undue assistance from third parties. No other than the resources and references indicated in this thesis have been used. I have marked both literal and accordingly adopted quotations as such. There were no additional persons involved in the intellectual preparation of the present thesis. I am aware that violations of this declaration may lead to subsequent withdrawal of the degree.

Dresden, November 20, 2018

Sebastian Bittrich